APLICAÇÃO DAS TÉCNICAS DE MINERAÇÃO DE TEXTOS E SISTEMAS ESPECIALISTAS NA LIQUIDAÇÃO DE PROCESSOS TRABALHISTAS

Antonio Alexandre Mello Ticom

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

.

Aprovada por:

_____
Prof. Nelson Francisco Favilla Ebecken, D.Sc

_____
Prof.ª Beatriz de Souza Leite Pires de Lima, D. Sc

_____
Prof.ª Sayonara Grillo Coutinho Leonardo da Silva, D.Sc.

_____
Prof.ª Valéria Menezes Bastos, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2007

TICOM, ANTONIO ALEXANDRE MELLO

Aplicação de Mineração de Textos e Sistemas Especialistas na Liquidação de Processos Trabalhistas Especialistas [Rio de Janeiro] 2007

VIII, 101 p. 29,7 cm (COPPE/UFRJ, M. Sc., Engenharia Civil, 2007)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de Textos
2. Categorização de textos
3. Sistemas Especialistas
4. Sentenças Trabalhistas

I. COPPE/UFRJ II. Título (série)

AGRADECIMENTOS

Aos meus pais pela educação e criação que me deram.

A minha esposa e filhos pela paciência durante este longo trabalho.

Aos Exmos. Juízes, Dr. Sergio da Costa Apolinário, Dr. Helio Ricardo Silva Monjardim, Dr. Paulo de Tarso Machado Brandão, Dra. Gisela Ávila Lutz e Dr. André Luis Amorim Franco, Dr. Carlos Eduardo Maudonet, Dr. Maurício Madeu, Dra. Maria Letícia Gonçalves, Dra. Alba Valéria Guedes Fernandes da Silva, por me apoiarem na área de Perícia Trabalhista.

A minha orientadora Prof.ª Beatriz de Souza Leite P. de Lima, por ter me herdado na orientação e dado o conhecimento necessário para desenvolver este trabalho.

Ao Marco Aurélio, Rodrigo Fernandes e Carlos Almeida pela ajuda no desenvolvimento do Sistema Especialista.

Aos professores Juan Lazo e Geraldo Xexeo pelo apoio.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APLICAÇÃO DE MINERAÇÃO DE TEXTOS E SISTEMAS ESPECIALISTAS NA LIQUIDAÇÃO DE PROCESSOS TRABALHISTAS

Antonio Alexandre Mello Ticom

Setembro / 2007

Orientadora: Beatriz de Souza Leite Pires de Lima

Programa: Engenharia Civil

A partir da evolução tecnológica dos meios de processamento de dados, principalmente a capacidade de processamento e armazenamento, surge uma nova área de pesquisa denominada “Extração de Conhecimento em Banco de Dados”. Dentre elas esta a Mineração sobre dados Não Estruturados (Text Mining) e Sistemas Especialistas. Este trabalho tem por objetivo apresentar os resultados da aplicação das Técnicas de Mineração de Dados em Textos Não Estruturados utilizando metodologias Probabilística, Linear por Ordenação e de Indução de Regras na Categorização de Textos, como também de Sistemas Especialistas, em Sentenças Judiciais da Área Trabalhista. O trabalho realizado procura informatizar, por completo, desde a fase em que o Juiz confere a sentença, relativo a uma reclamação trabalhista, passando pelas esferas judiciais seguintes (Embargos, Acórdãos, etc...) até o momento do cálculo final que a empresa reclamada deverá pagar ao empregado reclamante, contendo inclusive os valores a serem recolhidos de impostos (IR e INSS) aos cofres públicos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TEXT MINING AND EXPERT SYSTEM APPLIED IN LABOR LAWS

Antonio Alexandre Mello Ticom

September / 2007

Advisor: Beatriz de Souza Leite Pires de Lima

Department: Civil Engineering

Based on the technological evolution of data processing means, mainly the capacity of processing and storage, a new research field has emerged, called “Knowledge Discovery in Database”. Among these fields is Unstructured Data Mining (Text Mining) Specialist Systems. This work is aimed at presenting the results of the application of Unstructured Data Mining techniques using the following methodologies: Probabilistic, Linear Score and Rule Induction in the Categorization of Texts, as well as Specialist Systems, in Labor-related Judicial mandates. The work full automatize, since the Judge gives de sentence, related a work law, passing through the phases after the sentence until the moment for calculate the final value that the company have to pay to the employee, including the values it is supposed to be collected to the government.

Índice

1 Introdução

1.1 Introdução

A informatização dos meios produtivos, com o avanço da tecnologia, principalmente a velocidade de processamento e a redução do custo do armazenamento em meio magnético, tornou cada vez mais fácil e barato coletar, gerar e arquivar informações por meio das transações eletrônicas, dos novos equipamentos científicos e industriais para observação e controle como também dos dispositivos de armazenamento em massa. Conseqüentemente, as grandes empresas passaram a ter armazenado grande volume de informações.

Os recursos de análise de dados tradicionais são inviáveis para acompanhar esta evolução e este volume de dados. Ocorre também que o melhor aproveitamento das informações permite um ganho de competitividade em relação aos concorrentes. A solução encontrada foi então criar ferramentas de automatização das tarefas repetitivas e sistemáticas de análise de dados; ferramentas de auxílio para as tarefas cognitivas da análise e a integração destas ferramentas em sistemas inteligentes, apoiando o processo completo de descoberta de conhecimento para a tomada de decisão.

No início da década de 90, surge então uma área de pesquisa para a análise de grandes volumes de informações com objetivo de identificar a validade, a utilidade, o significado, o desconhecido e o inesperado do relacionamento entre os dados (KRUSE, 2003), denominado Descoberta do Conhecimento em Banco de Dados (Knowledge Discovery Database).

Dentro da Descoberta de Conhecimento em Banco de Dados, está inserida a Mineração de Dados, também conhecida como Data Mining, que consiste em um conjunto de técnicas e ferramentas para identificar padrões (conhecimentos) inseridos em grandes massas de dados (HAN, 2001).

Neste contexto, surgiu e vem sendo utilizada cada vez mais a área de Mineração de Texto (MT), definida por TAN (1999) como a extração de padrões interessantes e não triviais em textos, ou também a extração de conhecimento em documentos não estruturados. Os resultados destes trabalhos ajudam bastante na tomada de decisão e

tem-se mostrado de grande utilidade para a área denominada Inteligência de Negócios (*Business Intelligence*).

A MT possui diversas áreas de aplicação, dentre elas podemos citar: Classificação/Categorização, Clusterização, Sumarização, Indexação, Extração da Informação, Extração de Características, entre outras (LOPES, 2004). A aplicabilidade prática destes assuntos pode ser vista, como por exemplo, em: Classificação – na seleção de mensagens eletrônicas (e-*mails*) do tipo *spam* em uma conta de endereços eletrônicos; Extração da Informação – em *sites* de busca como o *Google.*

A maioria das aplicações na área de Descoberta de Conhecimento em Informações Não-estruturadas (MT) é composta de etapas (figura 1.1) executadas em determinada seqüência especifica, tal como: Obtenção, Preparação e Processamento dos dados.

A primeira etapa, denominada Obtenção ou Coleta dos Dados, tem por objetivo localizar as informações necessárias em sua forma mais bruta e capturá-las para posterior tratamento.

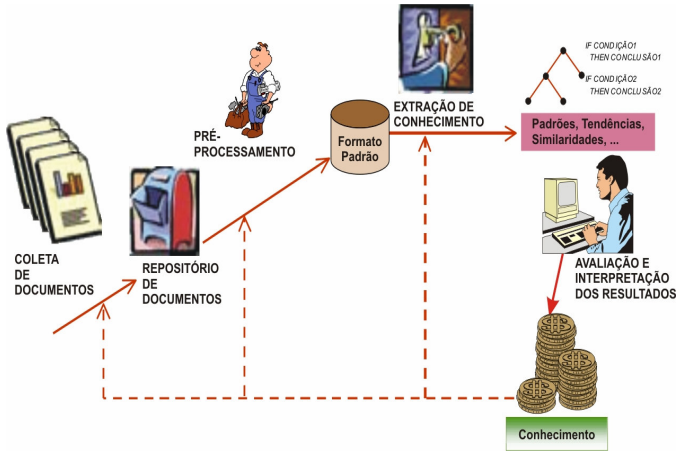


Figura 1.1 - As etapas de um processo de mineração de textos.

A fase seguinte, Preparação dos Dados, na maioria das vezes, é a mais trabalhosa e demorada. Esta etapa consiste de várias partes: conversão do texto para formato padronizado, normalmente XML (*eXtensible Markup Language*) (BRAY, 2000); conversão de todo o texto como minúsculo ou maiúsculo (*case folding)*, retirada de palavras desnecessárias (*stopwords*), redução das palavras ao menor radical (*stemming*) e redução de palavras por meio de dicionário de dados (*thesaurus*).

Na etapa seguinte os dados são processados. Na área de classificação/categorização, que é um dos principais objetivos deste trabalho, existem várias técnicas, tais como as Probabilísticas – Naive Bayes – (MCCALLUM, 1998), passando pelo popular *Support Vector Machine* – SVM – (JOACHIMS, 1998), as técnicas de Indução de Regras, Método Rocchio (ROCCHIO, 1971), "Vizinho mais Próximo", por Árvore de Decisão, dentre outras.

Finalmente, são apresentados os resultados por intermédio dos indicadores de medição de desempenho.

Em que pese utilizar grande volume de texto no dia-a-dia, a área jurídica foi uma das últimas a se informatizar, mas agora se depara com oportunidade de tornar-se grande usuária dessas metodologias que tratam grande volume de dados não estruturados.

Paralelamente, os processos judiciais trabalhistas específicos envolvem cálculos de valores devidos a funcionários e ex-funcionários. Nestes processos, essas pessoas reclamam uma ou mais verbas salariais, supostamente pagas de forma incorreta. Se o juiz deferir favoravelmente ao empregado, é necessária a utilização de várias regras para se apurar o valor correto.

O emprego de regras remete a grande oportunidade de se utilizar 'Sistemas Especialistas' (SE) na liquidação destes processos. A utilização de SE vem sendo cada vez mais utilizada na área jurídica. Um exemplo pode ser visto no teste feito com um grupo de advogados não especializados na área de direito autoral australiano realizado por O'CALLANGHAN et al. (2003).

1.2 Motivação e objetivo

Devido ao fato da área jurídica possuir grande volume de documentos e dados não estruturados, vislumbram-se muitas oportunidades de utilizar as técnicas de Mineração de Textos para extrair conhecimentos dos mesmos.

Somado ao fato de que, quando é necessário converter as decisões judiciais trabalhistas em valores financeiros a serem recebidos pelos funcionários, são utilizadas grandes quantidades de regras as quais se não identificadas e automatizadas acarretam morosidade nos cálculos e, principalmente, grande probabilidade de erros, abrindo oportunidade então para a aplicação de SE.

Logo, o principal objetivo deste trabalho é, inicialmente, utilizar as técnicas de mineração de textos para classificar os pedidos deferidos pelos juízes trabalhistas, tal como o apresentado por TICOM (2007). O outro objetivo é integrar automaticamente, provendo então estas informações de insumo para um SE que, com base em regras obtidas com especialistas, irá apurar precisamente o valor que a empresa deve ao empregado como também os valores a serem recolhidos à Receita Federal, de Imposto de Renda, e à Previdência Social de INSS.

O fato das ferramentas de MT e SE trabalharem integradas poderia ser classificado como um grande avanço na liquidação de processos trabalhistas, pois, atualmente, para se calcular o valor devido em um processo, é necessário ler e interpretar manualmente os documentos das sentenças dos Exmos. Juizes e digitar todo o resultado da sentença em planilhas eletrônicas para se obter os resultados finais.

Na área jurídica, alguns trabalhos têm sido desenvolvidos empregando-se Sistemas Inteligentes, ou ainda de MT. Porém, na área trabalhista, um SE para Liquidação de Sentenças Judiciais Trabalhistas utilizando as técnicas de MT é algo totalmente inovador em âmbito nacional e até mesmo no exterior, porque a Justiça Especializada Trabalhista (CLT – Consolidação das Leis do Trabalho) existe somente em poucos países, entre os quais o Brasil.

Portanto, as principais contribuições deste trabalho consistem em, primeiramente, aplicar as técnicas de classificação de documentos, oriundas da mineração de textos no ambiente jurídico-trabalhista, agilizar, como também reduzir a probabilidade de erros, por meio do processamento por um SE que apura os valores devidos por uma empresa a um empregado.

Do ponto de vista numérico, para se ter a noção do volume de pessoas e recursos envolvidos na área Jurídica, somente no Rio de Janeiro, apresenta-se a seguir o cenário que a envolve. O Tribunal Regional do Trabalho da primeira região, no município do Rio de Janeiro, possui em torno de 74 (setenta e quatro) varas do trabalho, com aproximadamente 3.600 mil funcionários, 400 juízes, 430 mil processos e oito mil advogados. Estes processos armazenados nas varas possuem valor estimado de pedido total da ordem de R$ 8 bilhões.

1.3 Organização dos capítulos

Este trabalho está dividido nos seguintes capítulos: o primeiro capítulo inicia-se com a Introdução, em que se descreve uma breve conceituação das técnicas de Mineração de Texto como também as de Sistemas Especialistas e suas aplicabilidades no contexto do objetivo deste trabalho.

No segundo capítulo, é descrita a Mineração de texto com detalhamento de suas teorias e técnicas.

No terceiro capítulo, são apresentadas as aplicações, técnicas existentes, principalmente as que demonstram como extrair informações e regras tendo em vista a *expertise* dos técnicos em cada assunto por meio de SE.

A seguir, no quarto capítulo, é explicada a origem de uma reclamação trabalhista, seu transcorrer, passando pelo papel dos advogados, os tipos de pedidos existentes, as peças jurídicas, desde a exordial até um Acórdão do Tribunal Superior de Trabalho e, principalmente, sua respectiva liquidação, objetivo principal deste trabalho.

No quinto capítulo, é descrito como o sistema opera, desde a etapa de obtenção dos dados, passando pela preparação das informações para escolha do melhor classificador. Na segunda parte deste capítulo, é apresentada a geração da interface para o SE e posterior processamento deste.

No sexto capítulo, é demonstrado o processamento do sistema para um Estudo de Caso e seus respectivos resultados com a classificação de documentos, aplicando as técnicas de *Naive Bayes*, Método Linear por Ordenação, Indução de Regras, empregando-se os aplicativos TMSK (WEISS, 2004) para as duas primeiras técnicas, RIKTEXT (WEISS, 2004) para Indução de Regras. Ao final, são apresentados os resultados gerados pelo SE com relação ao valor que a empresa deve pagar ao funcionário.

No sétimo e último capítulo, são apresentadas as conclusões e algumas sugestões de implementações a serem feitas em futuros trabalhos.

2 Mineração de Textos – Técnicas e Teorias

A Mineração de Textos (*Text Mining*) é um conjunto de técnicas e métodos utilizados para extrair conhecimento de dados não estruturados. Este trabalho visa a apresentar inicialmente a forma de obtenção dos dados para serem tratados. A seguir será detalhada a fase de preparação das informações para o processamento, que consiste em algumas técnicas como *case folding*, retirada de *stopwords*, *stemming,* entre outras. O trabalho também irá mostrar as várias métricas necessárias para utilização nestas aplicações. Entre as quais podemos citar a medida do cosseno, distância euclidiana, coeficiente de Pearson etc. Ao final, serão mostradas as várias tarefas existentes dentro da Mineração de Textos (MT), tais como Classificação, Clusterização, Sumarização, entre outros.

MT tem grande potencial para expandir o total de informação disponível, basta que as mesmas sejam analisadas e modeladas da melhor forma possível, transformando dados em conhecimento. MT veio depois da conhecida metodologia de *Data Mining*. Usualmente os dados tratados nas aplicações de *Data Mining* estão em formato de uma planilha/matriz de duas dimensões. Uma das dimensões apresenta as características, campos ou variáveis e a outra dimensão, apresenta as várias ocorrências ou também chamados de registros. Em contrapartida, os dados para MT são não estruturados (textos), ou seja, estão em um formato livre. Estima-se que 85% dos dados corporativos estão em um formato não estruturado. Acrescenta-se a este potencial o volume das informações disponíveis e cada vez mais crescentes na Internet.

2.1 Obtenção da informação

A primeira etapa numa aplicação que utilize MT é buscar extrair os dados necessários para que os mesmos sejam tratados. Devido ao fato deste tipo de aplicação, diferente de Mineração de Dados (*Data Mining*), tratar dados não estruturados (textos), serão requeridas técnicas mais complexas. Seguem abaixo duas das principais técnicas para captura da informação necessária a MT.

2.1.1 Recuperação da informação

Durante a última década, a quantidade de informação em formato de texto acessível eletronicamente cresceu exponencialmente. Isto se deve principalmente ao crescimento da Internet. As tecnologias baseadas na Internet exploraram a disponibilidade desta grande coleção de documentos para desenvolver os sistemas de Recuperação da Informação (RI). BELKIN (1992) apresenta um modelo para RI, conforme a figura 2.1. RI é normalmente o primeiro passo quando se deseja manusear dados textuais de uma grande coleção de documentos importantes. No caso de páginas indexadas da internet, potentes motores de pesquisa, tal como o *Google,* retornam uma lista ordenada de documentos para uma dada pesquisa do usuário. Existem duas estratégias básicas de pesquisa: pesquisa baseada em consulta e pesquisa baseada em documento.

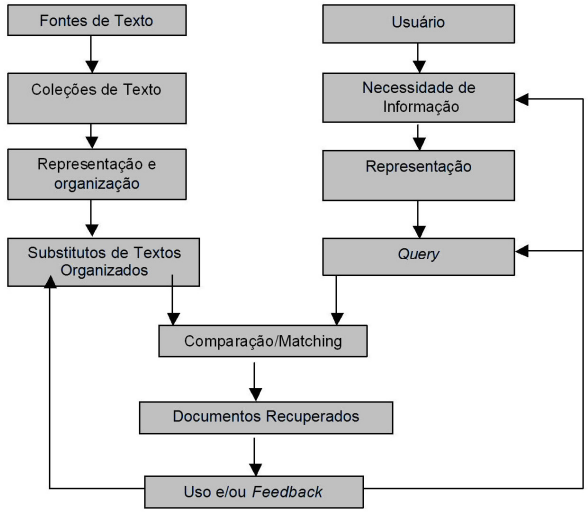


Figura 2.1 - Modelo de RI.

2.1.2 Filtragem da informação

A Filtragem da Informação (FI) tem recentemente atraído a atenção como um método de fornecer informação relevante. Os sistemas de FI cobrem uma grande variação de domínio, tecnologia e métodos, envolvendo o processo de entregar ao usuário a informação que ele deseja. A figura 2.2 mostra um modelo para FI. Os sistemas de FI se caracterizam por:

- ⇒ São aplicáveis em dados não e semi-estruturados (e-mails, documentos);
- ⇒ Manipulam um grande volume de dados;
- ⇒ Tratam principalmente com dados textuais;
- ⇒ São baseados no perfil do usuário;
- ⇒ Seu objetivo é remover os dados irrelevantes.

RI iniciou-se antes de FI. RI e FI são procedimentos similares porque ambos procuram obter informações sobre dados semi e não-estruturados. A grande diferença é que RI trabalha com consultas (*queries*) e FI com perfis (*profiles*) dos usuários.

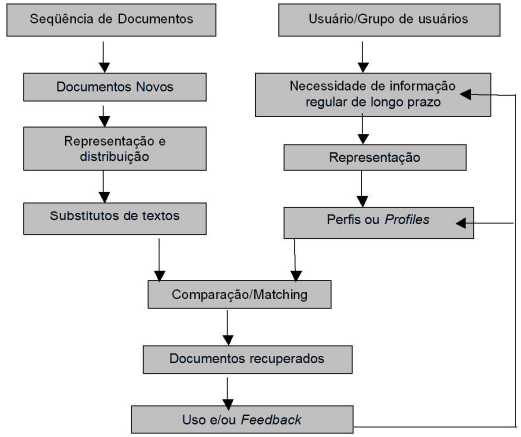


Figura 2.2 - Modelo de FI.

2.2 Preparação dos dados

Após a obtenção das informações desejadas, em uma aplicação de MT, estes dados que estão em um formato ainda bruto devem passar por alguns tipos de tratamentos com o intuito de prepará-los para posterior processamento. Seguem abaixo algumas das principais técnicas existentes.

2.2.1 Conversão de arquivo

Normalmente, os dados originais são convertidos para XML (*eXtensible Markup Language*), onde ficará mais fácil manipulá-los, visto que a estrutura desta linguagem é bastante adequada para tratar dados não estruturados. A linguagem XML é um formato que originalmente foi escrita para implementar estruturas de documento na Web (BRAY, 2000). Diferente de seu objetivo inicial, XML fez sucesso crescente como uma linguagem de representação de dados. A capacidade de representar qualquer tipo de dado como também ser uma linguagem padronizada mundialmente contribui muito para este sucesso. Igual a outras linguagens, XML tem regras e convenções que definem os elementos válidos. É uma linguagem que possui certos elementos (marcas) que podem ser utilizados para descrever estrutura e formato de partes do documento. O conjunto de elementos que podem ser usados no documento não é fixo, permitindo grande flexibilidade nos documentos XML e também bastante adaptabilidade a qualquer tipo de aplicação que a indústria requer. A linguagem XML é aberta, não foi desenhada por nenhuma grande corporação, mas sim por um consórcio (W3C) e tem por objetivo possibilitar uma linguagem altamente flexível e versátil. É uma linguagem simples, pois os documentos XML podem ser lidos pelos seres humanos e são de fácil entendimento. Portanto, é uma das ferramentas mais poderosas para representar textos a serem tratados por aplicações informatizadas. A figura 2.3 apresenta um trecho de arquivo em formato XML.

2.2.2 Transformação das letras (*Case Folding*)

Um dos primeiros tratamentos de dados a serem realizados é a transformação de todas as letras para maiúsculas ou minúsculas. Este procedimento pretende padronizar as palavras para que futuramente sejam identificadas no texto igualmente, com letras

maiúsculas ou minúsculas, possibilitando maior rapidez no processo de comparação entre caracteres.

<DOC>
<BODY>
HORAS EXTRAS
Afirma a Reclamante que desenvolvia trabalho, de segundas a sextas-feiras, no horário de 8h às 18h, do início do contrato a julho/1998, passando depois para prestação de serviços em dias alternados das 7h às 19h30 min, sempre sem intervalo para refeição, não recebendo pagamento por serviços extraordinários.
Defende-se o Reclamado informando inexistência de horas extras, afirmando jornada das 8h às 18h de segundas a quintas-feiras e das 8h às 17h nas sextas-feiras, com posterior alteração para escala de 12x36, das 7h às 19h, sempre com 1 hora de intervalo.
O Reclamado junta controles de horário, fls. 23/25, onde fica comprovado o horário alegado na defesa, quanto ao início e término de jornada, não havendo registro de intervalos, alegando que estes não precisam ser registrados, com invocação de norma administrativa indicada em defesa.
O horário de trabalho deve ser registrado, inclusive quanto aos intervalos, para fins de comprovação em juízo pelo empregador. Em não sendo acolhe-se o afirmado na inicial, quanto à inexistência de intervalos.
Registre-se que o Reclamado oferece defesa, no que respeita à jornada de trabalho, não invocando o instituto da compensação, sem comprovação de existência de contrato neste sentido.
A legislação estabelece como limite diário de trabalho 8h, sendo extras todas as horas trabalhadas em horário superior, com adicional de 50%, inexistindo previsão legal para a jornada de trabalho praticada pela Autora.
Assim, conforme prova nos autos, acolhe-se o horário indicado em defesa, como sendo de 8h às 18h, de segundas a quintas-férias e das 8h às 17h, nas sextas-feiras, até julho /1998, passando após, até o final do contrato, para 7h às 19h, em escala de 12x36, condenando-se a Reclamada no pagamento de horas extras, com adicional de 50%, sobre o trabalho prestado após a 8ª hora

diária, de segundas a sextas-feiras, com integração, por habituais, à remuneração de repousos semanais, 13º salários, férias com adicional de 1/3, aviso prévio e multa do art. 477, § 8º, da CLT.
Pela ausência de comprovação de intervalo de descanso e refeição de 1 hora, defere-se o adicional de hora extra de 50% ao dia, incidente sobre 1 hora de salário, em todos os dias de trabalho ao longo do contrato. É devido apenas o adicional, porque a hora normal já está paga, sendo utilizado o mesmo entendimento do Enunciado n. 85, do Colendo TST.
</BODY>
<TOPICS><TOPIC>hext</TOPIC></TOPICS>
</DOC>

Figura 2.3 – Texto no formato XML.

2.2.3 Retirada de palavras desnecessárias (*Stopwords*/*Stoplist*)

São palavras pouco úteis (*stopwords* ou *stoplist*) ou com baixo significado para tratamento em Mineração de Textos. São exemplos destas palavras artigos, preposições, conjunções, pronomes, tais como: de, assim, afim, agora, onde, outro, outros, ainda, a, o, que, vários, e, do, da, uns, em, um, para, é etc. O anexo A contém uma lista mais completa de *stopwords*. Normalmente, 40 a 50% do total de palavras de um texto são removidas com uma *stoplist* (KONGTHON, 2004; SALTON, 1983).

Ressalta-se que o processamento de textos, invariavelmente, trabalha com dimensionalidades muito grandes, o que requer grande espaço para armazenamento dos dados e alta capacidade de CPU. Portanto, é oportuno retirar as palavras que não agregam utilidade para a aplicação.

2.2.4 Redução ao menor radical de cada palavra (*Stemming*)

A última etapa da fase de Pré-processamento é a chamada *stemming*. Existem várias formas de trabalhar com *stemming*, cada uma com um propósito específico. Alguns algoritmos de *stemming* utilizam um dicionário, e outros trabalham com o sufixo das palavras. O *stemming* que trabalha com uma lista de sufixos tem como

finalidade reduzir as palavras, retirando seu sufixo, por meio de determinadas regras que dependem do idioma, até que a mesma fique com seu menor radical. Este processo tem como objetivo reduzir a quantidade de palavras diferentes no texto a serem tratadas. Dessa forma, reduz-se então a grande dimensionalidade das aplicações de MT, possibilitando utilizar menos espaço do computador e também menor tempo de execução de máquina. Ressalta-se que o objetivo do *stemming* não é chegar às regras básicas da lingüística do idioma, mas sim melhorar o desempenho das aplicações. Existem vários algoritmos de *stemming*. Dentre os mais empregados estão:

- **Método de Lovins -** Este método foi criado por LOVINS**,** em 1968. Consiste em um único passo de um algoritmo que remove aproximadamente 250 sufixos. É o mais agressivo dos quatro citados a seguir.

- **Método do Stemmer S -** É o mais simples e conservador, reduz apenas alguns poucos sufixos da língua inglesa "ies", "es" e "s".

- **Método de Porter -** É o mais popular atualmente. Sua intenção é remover sufixos com base em determinados critérios, sem se preocupar diretamente com os aspectos lingüísticos. Utiliza-se de várias fases para retirar em torno de 60 sufixos (Porter, 1980).

- **Stemming RSLP -** O Removedor de Sufixo da Língua Portuguesa – RSLP – (ORENGO, 2001) tem por objetivo ser um algoritmo de retirar sufixo para a língua portuguesa, que é baseado em regras, e cada uma delas pode ser expressa conforme a figura 2.4:

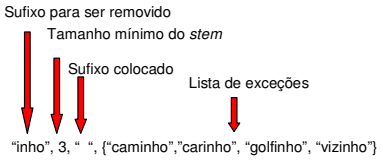


Figura 2.4 – Formato do RSLP.

O RSLP é composto de oito passos que precisam ser executados na ordem correta. A figura 2.5 apresenta a seqüência que os passos devem seguir. Cada passo tem um conjunto de regras, cada uma destas regras deve ser processada em determinada ordem e somente uma regra em cada passo pode ser aplicada. O sufixo mais longo possível é sempre removido primeiro, por causa da ordem das regras no passo. Por exemplo, o sufixo plural "es" deve ser testado antes do sufixo "s".

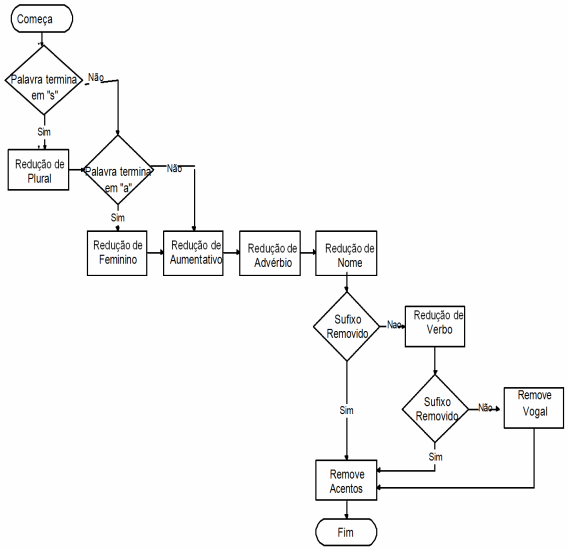


Figura 2.5 – Oito passos do RSLP.

Para uma melhor visualização, a figura 2.6 apresenta parte do arquivo após a retirada de *stopwords* e da execução do programa de *stemming*. Com o objetivo de fazer uma comparação, esta figura é a mesma apresentada na 2.3 antes de passar por estes procedimentos.

```
<DOC>
<BODY>
hor extr
afirm reclam desenvolv trabalho, segund sextas-feiras,
hor 8h 18h, inici contrat julho/1998, pass
prestac servic d altern 7h 19h30 min,
interval refeicao, receb pag servic
extraordinarios.
defende-s reclam inform inexist hor extras, afirm
jorn 8h 18h segund quintas-f 8h 17h
sextas-feiras, posteri alterac escal 12x36, 7h 19h,
1 hor intervalo.
reclam junt control horario, fls. 23/25, f comprov
hor aleg defesa, quant inici termin jornada,
hav registr intervalos, aleg precis
registrados, invocac norm administr indic defesa.
hor trabalh dev registrado, inclusiv quant intervalos,
fim comprovac juiz empregador. s acolhe-s
afirm inicial, quant inexist intervalos.
registre-s reclam oferec defesa, respeit jorn
trabalho, invoc institut compensacao, comprovac
exist contrat n sentido.
legislac estabelec limit diari trabalh 8h, s extr
tod hor trabalh hor superior, adic 50%,
inexist previs leg jorn trabalh pratic autora.
assim, conform prov autos, acolhe-s hor indic defesa,
s 8h 18h, segund quintas-fer 8h 17h,
sext feiras, julh /1998, pass apos, f contrato,
7h 19h, escal 12x36, condenando-s reclam pag
hor extras, adic 50%, sobr trabalh prest 8ª hor
diaria, segund sextas-feiras, integracao, habituais,
remunerac repous semanais, 13º salarios, ferias adic 1/3,
avis previ mult art. 477, § 8º, clt.
```



```
aus comprovac interval descans refeic 1 hora,
defere-s adic hor extr 50% dia, incid sobr 1 hor
salario, d trabalh long contrato. dev apen
adicional, porqu hor norm paga, s util
entend enunci n. 85, col tst.
</BODY>
<TOPICS><TOPIC>hext</TOPIC></TOPICS>
</DOC>
```

Figura 2.6 - Texto XML da figura 2.3 após retirada de *stopwords* e execução do algoritmo de *stemming* RSLP.

Ao mesmo tempo em que é uma poderosa ferramenta para melhorar o desempenho da aplicação de mineração de texto, se for mal trabalhado, os algoritmos de *stemming* podem prejudicar consideravelmente o resultado da análise. Os maiores riscos envolvidos neste processo são:

a) *Under-stemming* – quando um sufixo não é removido ou quando o algoritmo de *stemming* retirou um sufixo menor do que poderia;

b) *Over-Stemming* – é o contrário do anterior, quando o procedimento de *stemming* retirou mais sufixo do que deveria, ou seja, retirou parte do radical, acabando por gerar uma nova palavra sem relação com o texto como a anterior.

c) *Mis-stemming* – foi apresentado por Porter em adição ao *Under-stemming* e *Over-stemming* e significa quando o *stemming* tira parte da palavra, pois parecia um sufixo, mas não era.

Existem vários trabalhos apresentando o efeito do *stemming* no desempenho de aplicações de *Text Mining*. KRAAIJ (1996) fez uma pesquisa de revisão de *stemming* e identificou que vários fatores afetam seu resultado, tais como: a lingüística da língua, o tamanho do documento, entre outros.

2.2.5 Dicionário de dados (*Thesaurus*)

Uma boa alternativa para melhorar os resultados de uma aplicação é utilizar um dicionário de dados que correlacionam palavras diferentes e comuns a uma única palavra em todo o texto, ou seja, montar uma relação de várias palavras para uma única palavra que possa substituí-las sem alterar o contexto. Como exemplo, podemos citar as palavras "rua", "avenida", "estrada", que poderiam ser associadas a uma única palavra que é "rua". Um outro exemplo que se relaciona com uma aplicação jurídica é quando as palavras "defere", "deferido", "procedente", "procede", poderiam ser padronizadas como "deferido".

2.3 Medidas de avaliação

Nos procedimentos de Mineração de Textos (MT), sempre são utilizadas medidas matemáticas. Estas podem servir para avaliar a distância entre dois vetores, ou ainda quando se deseja atribuir pesos às palavras mais relevantes de um texto, e principalmente na mensuração do desempenho das técnicas de MT, tais como: classificação, clusterização, extração de características, entre outras. Por isso, antes de apresentar as áreas de aplicações da MT, será mostrado a seguir algumas das principais medidas de avaliação existentes.

2.3.1 Medidas de similaridade

Existem várias técnicas estatísticas e matemáticas para avaliar semelhança. As aplicações de MT utilizam métodos numéricos para identificar a similaridade entre os documentos ou entre estes documentos e as consultas. Citamos a seguir algumas das principais medidas existentes:

2.3.1.1 Medida de similaridade do cosseno

Tem grande utilização em medidas de documentos. Se existirem dois vetores, a medida do cosseno entre estes dois vetores será um menos o cosseno do ângulo formado

entre eles. A medida do cosseno será grande (perto de um) se os vetores forem quase ortogonais (este caso significa que existem poucas palavras comuns entre os documentos), e pequena (perto de zero) se os vetores forem similares (grande quantidade de palavras comuns a ambos). A expressão do cosseno para avaliar a similaridade entre dois documentos pode ser escrita pelas equações 2.1 e 2.2 (FULLAM, 2002):

$$M_{Cos} = \frac{\sum_{k=1}^{j} (d1_k \bullet d2_k)}{\sqrt{v_{d1} \bullet v_{d2}}} \qquad (2.1)$$

Onde:

d1 e d2 são documentos representados por vetores

j é igual ao total de termos

• representa produto escalar

$$v_{d1} = \sum_{k=1}^{j} d1_k^{\,2} \qquad (2.2)$$

2.3.1.2 Distância Euclidiana

Uma das medidas de distância (equação 2.3) mais popular para características contínuas é a Euclidiana (JAIN, 1999), em que pese não trazer bons resultados quando utilizada com documentos.

$$D_{Euc} = (\sum_{k=1}^{j} (d1_k - d2_k)^2)^{1/2} \qquad (2.3)$$

2.3.1.3 Coeficiente de correlação de Pearson

Dadas duas amostras de observações medidas em uma escala de intervalos ou razões, podemos medir o grau de associação linear entre elas por intermédio do coeficiente de correlação de Pearson ou simplesmente coeficiente de correlação amostral. Assumindo que ambas variáveis (X e Y) são intervalos entre variáveis, as mesmas são bem aproximadas por uma distribuição normal como também sua distribuição conjunta é normal bivariada. O coeficiente de Pearson (BOLBOACĂ, 2006) é dado pela expressão 2.4:

$$C_{Pea} = \frac{\sum_{k=1}^{j} (d1_k - \overline{d1})(d2_k - \overline{d2})}{\sqrt{(\sum_{k=1}^{j}(d1_k - \overline{d1})^2 (\sum_{k=1}^{j}(d2_k - \overline{d2})^2)}} \qquad (2.4)$$

Onde $\overline{d1}$ e $\overline{d2}$ são iguais à média da amostra de d1 e d2.

Este coeficiente de correlação pode variar entre -1 e 1. Ele assume o valor 1 quando os pontos estão exatamente sobre uma reta em declive positivo. Neste caso, um aumento em uma das variáveis corresponde necessariamente a um aumento na outra. R assume o valor -1 quando os pontos estão exatamente sobre uma reta de declive negativo. Nesta situação, um aumento em uma das variáveis corresponde a uma diminuição na outra. Estes dois casos correspondem ao máximo de associação linear, que é possível observar entre duas amostras. Quando as amostras são independentes, o valor do coeficiente será próximo de zero ou mesmo zero. Uma interpretação usual do coeficiente de correlação amostral passa por considerar o seu valor elevado ao quadrado, R^2, a que se chama coeficiente de determinação. Uma vez que $-1 \leq R \leq 1$, o coeficiente de determinação está sempre entre 0 e 1. Resumindo, o coeficiente de correlação de Pearson mede o grau de associação linear entre duas variáveis medidas em uma escala de intervalos ou razões. Se as variáveis tiverem distribuição Normal podemos efetuar um teste de hipóteses para averiguar se o coeficiente de correlação da população é significativamente diferente de zero, o que significará, nesse contexto, que as variáveis são independentes. Convém sempre construir um diagrama de dispersão para ter uma idéia sobre a linearidade da relação entre as variáveis.

2.3.1.4 Coeficiente de correlação de Spearman

O coeficiente de Spearman é normalmente utilizado quando não se pode utilizar o coeficiente de Pearson, ou seja, quando não se podem garantir os pressupostos da realização do teste de hipótese a este coeficiente, se houver duas variáveis medidas

apenas em uma escala ordinal, ou ainda se apresentarem uma relação não linear, mas monótona (se uma aumenta a outra tem sempre tendência a aumentar ou a diminuir).

A fórmula simplificada para calcular o *rank* de Spearman (BOLBOACĂ, 2006) é dada pela equação 2.5:

$$C_{Spm} = 1 - \frac{6\sum Di^2}{q(q^2 - 1)} \qquad (2.5)$$

Onde *Di* é a diferença entre cada par do *rank* $d1_k, d2_k$ e *q* é igual à quantidade da amostra.

Embora um coeficiente nulo não implique independência total, este teste é utilizado na prática para averiguar se a associação entre variáveis é significativa ou não, entendendo-se por associação uma correlação não nula.

2.3.2 Atribuição de pesos (*weighting*)

As aplicações de MT, para se fazer boas previsões, utilizam vetores com uma dimensionalidade muito grande de palavras/características. Para diferenciar as características mais relevantes, utiliza-se a atribuição de pesos. Os três pesos mais utilizados estão descritos a seguir:

- **Binário -** O valor unitário (*true*) é atribuído a um termo *t* quando o mesmo é encontrado no documento *d* e zero (*false*) quando não encontrado. Esta representação é muito simples e deve ser utilizada dependendo do domínio. Normalmente, utilizam-se medidas estatísticas levando em consideração a freqüência dos termos na coleção de documentos, tal como será descrito nos dois próximos itens.

- **TF –*Term Frequency* – (Salton, 1983) -** É definida como o número de vezes que o termo *t* é encontrado no documento *d*. Quando termos com alta freqüência aparecem na maioria dos documentos da coleção, os mesmos

passam a não fornecer informação relevante para a diferenciação dos documentos.

- **TF*IDF (*Term Frequency – Inverse Document Frequency*) -** A medida IDF é definida como o *log nd/t*

Onde *nd* é igual ao número de documentos em que o termo *t* é encontrado pelo menos uma vez.

Esta medida favorece termos que aparecem em poucos documentos de uma coleção. Logo, é possível trabalhar com um novo indicador juntando as medidas TF*IDF. Pode também ser utilizado um fator de normalização para fazer com que documentos de tamanhos diferentes possam ser tratados com a mesma importância.

2.3.3 Medidas de desempenho

São indicadores utilizados para avaliar o desempenho das técnicas de Mineração de Textos, como, por exemplo, para medir o resultado de uma rotina de classificação de textos, de uma clusterização de documentos, entre outros. São normalmente expressos em percentual, e, quanto maior o resultado, indicam uma melhor performance. Dentre os principais, podemos citar:

- **Índice de Precisão (*Precision*) -** Medida analisada no âmbito de cada classe. É a razão entre o número de documentos corretamente classificados e o número total de documentos associados à classe. Pode ser definida pela expressão 2.6:

$$\frac{x}{x+z} * 100 \quad \text{- valor em percentual \%} \qquad (2.6)$$

Onde

x = número de documentos associados a uma classe c e corretamente classificados como pertencente a esta classe.

z = número de documentos não associados à classe c, mas classificados como pertencentes a esta mesma classe.

- **Cobertura (*Recall*) -** É a razão entre o número de previsões corretas positivas sobre o número de documentos da classe positivos. A seguir, a expressão 2.7 que define cobertura:

$$\frac{x}{x+y} * 100 \quad \text{- valor em percentual \%} \qquad (2.7)$$

Onde x tem a mesma definição utilizada na Precisão e y é o número de documentos associados a classe c e não classificados.

A figura 2.7 (JIZBA, 2000) representa a definição de precisão e cobertura.

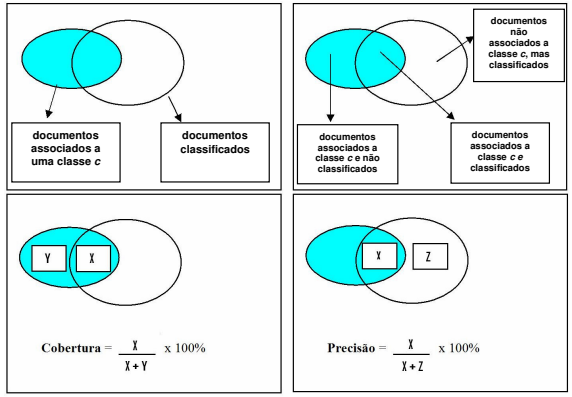


Figura 2.7 - Demonstração do cálculo da precisão e cobertura (JIZBA, 2000).

- **Medida F *(F-measure)*** - A medida F pode ser definida em função da precisão e da cobertura e é dada pela expressão 2.8:

$$\text{Medida F} = \frac{2}{1/\,precisao + 1/\,cobertura} \qquad (2.8)$$

2.4 Tarefas de mineração de textos (MT)

Nos itens anteriores foram apresentadas as formas de obtenção e preparação dos dados, e de algumas medidas de avaliação. Agora, serão apresentadas as tarefas de MT, que incluem desde a extração de características, passando pela classificação até a clusterização de documentos. Todas utilizam informações não estruturadas (textos). Como exemplos destas aplicações na prática, podem ser citados: uma ferramenta de busca como o *Google*; um filtro para *spam* em uma conta de *e-mail* e uma implementação feita por TICOM (2007) para classificar pedidos em sentenças judiciais trabalhistas. Existem vários tipos de técnicas que podem ser utilizadas em cada uma destas aplicações. Segue abaixo as principais áreas de aplicações e pesquisas de MT.

2.4.1 Sumarização

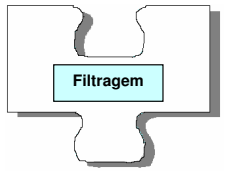
A sumarização tem a finalidade de extrair as informações mais representativas, normalmente palavras ou sentenças, do documento, que poderá ser lido pelo usuário, em vez do documento original, visto que o significado de ambos deve ser o mesmo (RADEV et al, 2001). Quando a entrada consiste em mais de um documento, denominamos que a sumarização é "multi-documento". A sumarização se baseia no princípio da redundância e na distribuição desigual de informações.

Devido a cada vez mais crescente quantidade de informações nos últimos tempos, principalmente na Internet, a demanda pelas técnicas de sumarização de textos começa a aparecer para as empresas comerciais. Estas começam a utilizar cada vez mais as ferramentas de recuperação das informações e os sistemas de banco de dados. Este desenvolvimento oferece oportunidade para desafios em pesquisas em sumarização de textos. O desenvolvimento cria uma dependência dos sistemas de sumarização quando é necessário tratar grandes volumes de texto.

2.4.2 Extração de informações

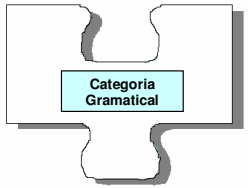
Os sistemas de Extração de Informações (EI) processam documentos com linguagem natural e identificam termos específicos relevantes. Estes termos podem ser utilizados para apenas separar o texto nas partes mais interessantes ou para preencher um formulário/arquivo eletrônico com os campos pré-definidos. EI pode ser útil em vários

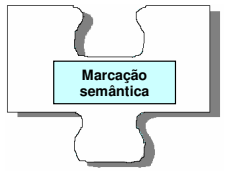
segmentos, tais como: na medicina, área jurídica, eletrônica, engenharia, entre outros. A maioria das pesquisas em linguagem natural empregam técnicas estatísticas e se baseiam em um contexto muito limitado ou em técnicas simbólicas, como árvores de decisão. Os sistemas de EI normalmente utilizam programas indutores de lógica, que consistem em pesquisas por padrão do tipo específico para o geral (*bottom-up*) e são caracterizados pelo preenchimento de modelos. A figura 2.8 apresenta os componentes de um sistema típico de EI.



Nível de texto – Determina a relevância do texto ou parte do texto baseado na estatística das palavras ou na ocorrência de padrões específicos do texto.

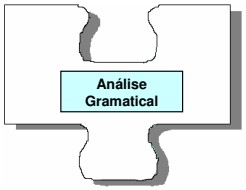
Nível de palavra – marca palavras de um texto de acordo com sua categoria gramatical Usualmente utiliza métodos estatísticos treinados por um texto pré-marcado.





Nível de frase – reconhece a maioria dos tipos de frases no domínio e as marca com informações semânticas.

Nível de sentença – Mapeia os elementos da sentença numa estrutura que mostra a relação entre eles.



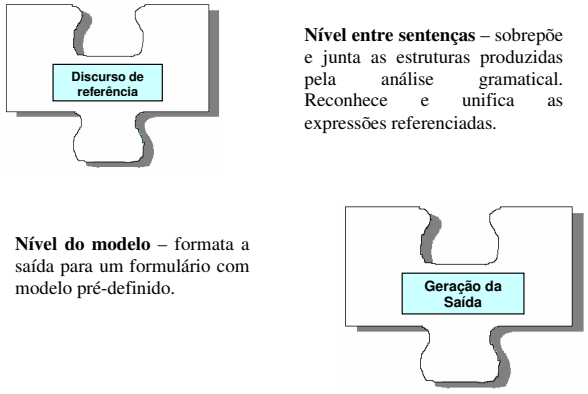


Figura 2.8 - Componentes de um sistema típico de EI (COWIE, 1996).

2.4.3 Extração de características

Extração de Características (EC) consiste em extrair termos relevantes para a aplicação segundo objetivos pré-definidos do texto, tal como buscar todos os nomes próprios em um texto específico; trocar a cor de todos os nomes de cidade e nomes de empresa. Para um exemplo de uma aplicação de EC, pode ser desenvolvida uma rotina que busca na WEB os *sites* de bancos, mostrando se uma tarifa de um produto/serviço foi alterada em relação à última capturada de cada um dos bancos. Isto permite a um determinado banco avaliar se sua tarifa está alta ou não em relação ao mercado. Da mesma forma, pesquisas de cotação de preço podem ser feitas utilizando as técnicas de EC. Também pode ser definida como uma subárea da Extração de Informação, especificamente com o objetivo de extrair características desejadas do texto, em vez de informações como um todo. As informações nas empresas estão originalmente em um formato não estruturado, logo, de difícil tratamento para o processamento automático dos sistemas tradicionais, por isso, utiliza-se as técnicas de EC. As informações geradas pelos sistemas de EC são muito úteis para a área de Inteligência de Negócios. Um procedimento de EC é normalmente decomposto em uma seqüência de passos de processamento, em que estão incluídos: *tokenização*, segmentação de sentenças, nome

das entidades a serem identificadas, interpretação semântica, preenchimento de modelo e junção.

2.4.4 Indexação

Indexação é a tarefa da MT que trata da identificação dos termos mais representativos, normalmente os que são mais utilizados, existentes em uma coleção de documentos e a posterior disponibilização destes termos em meios magnéticos para acesso rápido com objetivo de aumentar a rapidez nas consultas de determinadas aplicações. Esta área é bastante similar à área de indexação em banco de dados tradicional.

SALTON (1989) apresenta um sistema automático de indexação que contém as funções típicas, tal como: dicionário de dados, *stopword*, *stemming* e os termos para formação das frases. Em um primeiro momento, o algoritmo identifica as palavras individualmente. A seguir, uma lista de *stopwords* é utilizada para remover as palavras não significativas. Depois desse passo, é executada uma rotina de *stemming* para reduzir as palavras a seu menor radical. Ao final, são formadas as frases combinando as palavras adjacentes. A indexação automática de textos identifica nos textos os termos mais usados por diferentes grupos de pessoas. No Modelo do Espaço Vetorial de SALTON (1989) é identificado um peso para cada termo com o objetivo de medir a importância. Dentre várias técnicas disponíveis, as mais utilizadas são: binária, freqüência do termo (tf) e freqüência do termo/freqüência inversa do documento (tf*idf), apresentada anteriormente nesta dissertação.

Uma variação do sistema apresentado por SALTON, chamada de Indexação Semântica Latente, é descrita por DEERWESTER (1990). Este método se propõe a ultrapassar a deficiência na pesquisa do termo, baseado no tratamento da falta de confiança dos dados associados aos termos do documento como um problema estatístico. Interpreta-se como existindo um nível inferior de estrutura semântica latente nos dados que não é visto devido à característica aleatória da palavra escolhida em relação à função de recuperação de dados. São usadas então técnicas para estimar esta estrutura latente e melhorar estes pontos obscuros.

2.4.5 Clusterização ou agrupamento

Clusterização é o agrupamento dos documentos de uma coleção em N grupos, com a maior semelhança possível, baseado em uma métrica pré-definida. Assim como a maioria das áreas de aplicação da MT, os métodos de Clusterização também requerem uma fase de pré-processamento dos dados. Nesta fase, são realizadas transformações das letras para minúscula (*case folding*), a retirada de termos desnecessários (*stopwords*), redução da palavra ao menor radical (*stemming*), que serão descritas com maior detalhe em capitulo posterior. A maioria dos algoritmos de *clustering* de texto se baseia no Modelo do Espaço Vetorial (SALTON, 1989), no qual cada documento é representado como um vetor de freqüências de *t* termos, como demonstra a equação 2.9:

$$D = (TF_1,.....,TF_t) \qquad (2.9)$$

Onde TF é igual à freqüência de cada termo, detalhada em capítulo anterior.

Normalmente, o próximo passo é a normalização dos vetores para possibilitar fazer comparações com documentos de tamanhos diferentes. Este modelo acarreta vetores com uma alta dimensionalidade. Ao final, para comparar a semelhança entre dois documentos *d* do modelo do espaço vetorial, alguma métrica é utilizada. Uma das mais freqüentes é a do cosseno, que mede o ângulo entre dois vetores e foi descrita no item 2.3.1.1.

Os algoritmos padrões de *clusterização* são normalmente divididos em algoritmos particionados como o *k-means* e o *k-medoid* ou algoritmos hierárquicos do tipo do *single-link* ou *average-link*. Um estudo (STEINBACH, 2000) foi feito comparando os algoritmos particionados com os hierárquicos. O resultado demonstrou que o *k-means* obteve uma melhor eficiência como também uma melhor qualidade do *cluster*.

Um trabalho interessante foi desenvolvido por Xavier (2005) no qual o problema da clusterização pode ser resolvido pelo método *Smoothing Hyperbolic*.

2.4.6 Classificação

A área de aplicação denominada como Classificação tem por objetivo identificar, por semelhança, cada novo documento como um dos tipos de categorias (classes) previamente definidas. Esta área começou a ser utilizada nos anos 60 do último século**,** quando era utilizada para aliviar os serviços dos indexadores científicos de literatura. Somente na década de 1990 a Classificação de Textos começou a crescer, devido à necessidade de tratar o crescente número de documentos de texto em meio magnético.

Atualmente, é utilizada em diferentes aplicações, tais como: personalização de informações para entrega; filtrando conteúdos indesejáveis; identificando padrões; classificando as páginas da Internet em um catálogo hierárquico; diagnósticos médicos; geração automática de meta-dados; detectando fraudes; aprendizado de ontologias; entre outras. Existe uma grande quantidade de métodos e técnicas que podem ser aplicados para classificação de documentos. A principal divisão entre os métodos existentes são os lineares e os não lineares. Comparando com outros métodos, os classificadores lineares são simples e têm um modelo de treinamento muito mais fácil de ser interpretado. Também demonstram ser muito efetivos e seu desempenho apresenta-se como um dos melhores para categorização de textos. Segue abaixo um resumo dos principais métodos existentes.

2.4.6.1 Naive Bayes

O classificador linear Naive Bayes (MCCALLUM & NIGAM, 1998) é bastante utilizado na comunidade de MT, especialmente para as aplicações de Classificação de Textos. É um método probabilístico, no qual se assume que todas as variáveis são independentes da variável de classificação, o que o torna muito fácil para criar uma rede estruturada e não obriga a geração de um algoritmo de aprendizado. Este classificador se baseia no teorema de Bayes com a simplificação de que, após o treinamento, pode ser assumido que as características são independentes para uma dada classe.

Dado que o vetor de características é $D = (t_1,.....,t_n)$ e C a classe, a equação 2.10 (Rish, 2001) apresenta o cálculo da probabilidade.

$$P(D \mid C) = \prod_{k=1}^{j} P(t_k \mid C) \qquad (2.10)$$

2.4.6.2 *Support Vector Machine* (SVM)

Um dos mais populares classificadores do tipo linear. O SVM implementa a idéia de que seja construído um hiperplano com base no mapeamento dos vetores de entrada em um espaço de características com uma grande quantidade de dimensões. Quando os dados do arquivo de treino são separáveis, a taxa de erro para o SVM pode ser definida pela equação 2.11:

$$h = R^2 / M^2 \qquad (2.11)$$

Onde R é o raio da menor esfera que contém os dados de treinamento.

M é a margem que significa a distância entre o hiperplano e o vetor de treino mais perto do espaço de características.

Existem dois autores bastante conhecidos que desenvolveram grandes trabalhos nesta área de classificação utilizando SVM. O primeiro deles é Vapnik, um dos grandes nomes também do Aprendizado de Máquina e da Inferência Estatística, que muito contribuiu quando escreveu o livro *The Nature of Statistical Learning Theory* (VAPNIK, 1999), inicialmente em 1995, e a segunda edição em 1999. O segundo é Joachims, que criou uma variação do SVM, denominada *light* (JOACHIMS, 2002), livre e disponível no *site* http://svmlight.joachims.org. Joachims (1998) cita em seu artigo que foi VAPNIK (1999), na primeira edição em 1995, que fundamentou como o treinamento do SVM para o problema de reconhecimento de padrões pode ser resolvido por intermédio da otimização de uma função quadrática.

2.4.6.3 Regressão linear

Este método procura identificar uma função linear em que os dados de treinamento se enquadrem (ZHANG, 2003). O algoritmo *Linear Least Square Fit* (LLSF) é o método mais utilizado para estimativa de regressão linear (equação 2.12) e se equivale ao *Maximum Likehood Estimation*, quando *y* é influenciado pelo ruído Gaussiano.

$$f(z) = w^R z \qquad (2.12)$$

O algoritmo LLSF calcula um vetor de peso *w* baseado na minimização da perda quadrada entre o modelo de saída $w^R z$ e $f(z)$.

2.4.6.4 Regressão logística

A Regressão Logística (ZHANG, 2003) é bastante utilizada na estatística há um longo tempo, mas somente começou a ser aplicada no aprendizado de máquina recentemente, devido à próxima relação com o SVM. Embora não tenha sido tão utilizada até agora como o SVM e o LLSF, tem sido usada na classificação de textos e comparada com outros métodos de classificação linear, devido a sua performance ser comparável ao SVM. A regressão logística tenta modelar a probabilidade condicional *p(u|z)*. Para uma classificação na qual somente existam duas classes (binária), esta probabilidade pode ser modelada por meio da equação 2.13:

$$p(u \mid z, w) = \frac{1}{1 + \exp(-uw^R z)} \qquad (2.13)$$

Onde *p(u|z)* é a probabilidade condicional e $uw^R z$ a função

2.4.6.5 Método linear por ordenação (*Scoring*)

O método linear por ordenação (WEISS, 2004), em função de utilizar uma função linear com pesos para as características e um *bias*, é muito utilizado para tratar os problemas de classificação/categorização, visto que estes requerem uma capacidade de selecionar as características mais relevantes dentre um volume muito grande. Este método também é muito simples, dado que basta identificar as características mais importantes e deixar o algoritmo calcular um peso para cada uma delas. A equação 2.14 define o cálculo do *Scoring*.

$$Scoring(D) = \sum_j p_j l_j + b = pl + b \qquad (2.14)$$

Onde D é o documento, p_j é o peso da j-ésima palavra do dicionário, b uma constante e l_j é um ou zero, dependendo se a j-ésima palavra existia ou não no documento.

Comparação entre os métodos lineares descritos acima

A título de ilustração dos métodos descritos anteriormente, a seguir é apresentado uma comparação entre o desempenho dos principais classificadores lineares ZHANG (2001). Nesta comparação são empregados os seguintes classificadores lineares: *Linear Least Square Fit*, *Modified Least Square Least*, *Logistic Regression*, *Support Vector Machine*, *Modified SVM* e *Naïve Bayes*. Os resultados contemplam a utilização de 118 classes da base *Reuters* e também 36 classes da base de dados AS400 do call center dos clientes da *IBM*, e estão demonstrados nas (tabela 2.1) e (tabela 2.2) respectivamente.

	Naive Bayes	*Lin Reg*	*Mod Least Squares*	*Logistic Reg*	*SVM*	*Mod SVM*
Precisáo	77,0	87,1	89,2	88,0	89,2	89,4
Cobertura	76,9	84,9	85,3	84,9	84,3	83,7
Medida F	77,0	86,0	87,2	86,4	86,5	86,5

Tabela 2.1 – Resultados da base *Reuters*.

	Naive Bayes	*Lin Reg*	*Mod Least Squares*	*Logistic Reg*	*SVM*	*Mod SVM*
Precisáo	66,1	78,5	77,7	76,3	78,9	78,7
Cobertura	74,9	64,0	70,9	74,1	63,8	63,6
Medida F	70,2	70,5	74,1	73,8	70,6	70,4

Tabela 2.2 – Resultados da base de dados AS400 do *call center* dos clientes da IBM.

2.4.6.6 Indução de regras

O classificador por Indução de Regras tem como finalidade procurar palavras-chave no texto que permitam recuperar exatamente estes documentos, ou seja, encontrar uma ou mais palavras que servem para identificar univocamente um documento. A

grande vantagem deste método é a fácil compreensão da visualização dos resultados. Em contrapartida, o procedimento de achar as regras pode ser mais trabalhoso do que outros métodos, principalmente se estiverem sendo tratados grandes coleções de documentos e palavras gerando uma grande quantidade de regras. Para aperfeiçoar/facilitar a geração de regras, existem alguns algoritmos que aumentam o desempenho, entre os quais o Adaboost, descrito por (SCHAPIRE, 2001).

2.4.6.7 K-Vizinho mais próximo

O algoritmo K-vizinho mais próximo (KNN) utiliza uma técnica de classificação não-paramétrica, que se tem mostrado bastante eficaz em aplicações para reconhecimento de padrões. Esta técnica possibilita obter grande precisão na classificação em que os problemas têm uma distribuição desconhecida. Em contrapartida, as implementações tradicionais desta técnica tratam uma grande quantidade de vetores, acarretando uma alta complexidade computacional para o classificador. Portanto, a mesma torna-se lenta como também requer um grande espaço de memória do computador.

Para comprovar a lentidão e a necessidade de alto volume de espaço em disco pelas aplicações que implementam o KNN, foi utilizada a versão de uma aplicação, onde se usou algumas métricas de distância, entre as quais: euclidiana, manhattan, camberra e minimax.

Devido a estes problemas citados, várias otimizações têm sido desenvolvidas para melhorar este tipo de classificador. Uma delas foi desenvolvida por RAHAL (2004) e se baseia na tecnologia denominada *P-Tree*. Este formato utiliza o armazenamento dos dados em uma árvore, e as estruturas de dados numéricas são comprimidas e convertidas para binário. Esta forma de armazenamento de dados possibilita armazenar grande quantidade de informações e facilita os processos de mineração. De uma forma resumida, inicialmente, o algoritmo de classificação cria a matriz de termo por documento com a métrica TF*IDF. Esta matriz depois é convertida para o formato *P-tree*. A seguir, o algoritmo procura os k-vizinhos mais próximos. A fase de seleção esta descrita a seguir.

Após criar e ordenar os termos das *P-trees* de acordo com os valores do novo documento, o algoritmo, seqüencialmente e para cada termo da *P-tree (Pt)*, procura confirmar que o contador da raiz é maior ou igual a *k*. Este processo de reconstruir *Pt* é

repetido até que o resultado da nova *Pt* tenha o contador da raiz maior do que *k*. Depois da repetição com todos os termos da *P-tree*, o documento estará como o mais próximo do novo documento. A seguir, o algoritmo procura o rótulo da classe do novo documento. Posteriormente, para cada documento vizinho, é dado um peso baseado na sua similaridade. Depois, para cada rótulo de classe, é feita uma repetição para todos os termos do novo documento, calculando o número de vizinhos mais próximos que têm o mesmo valor deste termo para todos os termos do novo documento. KHAN (2002) demonstra que este algoritmo é mais preciso que o tradicional k-vizinho mais próximo.

2.4.6.8 Árvore de Decisão

Um classificador de texto do tipo árvore de decisão (MITCHELL, 1997) é uma árvore em que os nós internos são rotulados pelos termos, os ramos que partem dos nós são definidos pelos testes, levando-se em consideração o peso que o termo tem no teste do documento e as folhas pelas categorias. A maioria dos classificadores utiliza a forma binária para representar os documentos gerando conseqüentemente uma árvore binária. Existem vários pacotes para aprendizado por árvore de decisão, e a maioria das abordagens de árvore de decisão para Classificação de Textos utilizou um destes pacotes. Os mais populares são: ID3 (FUHR, 1991), C4.5 (COHEN, 1998), e C5 (LI, 1998).

2.4.6.9 Redes Neurais

O classificador de textos que utiliza redes neurais pode ser definido como uma rede de unidades onde as unidades de entrada representam os termos, as unidades de saída significam as categorias de interesse e os pesos nas conexões representam as relações de dependências. O mais simples tipo de classificador de rede neural é o *perceptron* (Dagan, 1997), que pode ser definido como um classificador linear.

2.4.6.10 Algoritmos *On-line*

Existem alguns algoritmos de classificação denominados *on-line*, que são caracterizados por permitirem que a previsão seja feita também por meio do "aconselhamento" (atribuição de pesos aos termos) dado por N especialistas.

Tipicamente, este tipo de algoritmo utiliza uma combinação de peso das previsões dadas pelos especialistas. Dois destes algoritmos (*Sleeping-Expets for Phrases* e *RIPPER*) estão descritos em COHEN (1999).

3 Sistemas Especialistas – Teoria e Técnicas

Durante as três últimas décadas, pesquisadores de inteligência artificial (IA) foram aprendendo a apreciar o valor do conhecimento específico do domínio como um requisito indispensável na resolução de problemas complexos (DOYLE, 1996). Os avanços em *hardware*, tecnologia de *software* e ciência cognitiva possibilitaram a construção de ferramentas e técnicas baseadas em conhecimento. Os sistemas baseados em conhecimento (SBC) fazem parte desta geração de técnicas e ferramentas.

Os sistemas especialistas constituem uma área da Inteligência Artificial. O objetivo de um sistema especialista (SE) é captar o conhecimento amplo de um especialista em uma determinada área, representar esse conhecimento em uma base e permitir ao usuário obter respostas a perguntas relacionadas à base de conhecimento do sistema.

Os SE fornecem conclusões acerca de assuntos especializados, por meio da emulação do raciocínio de um ou vários especialistas, em um domínio específico, ou seja, são sistemas com um conhecimento específico profundo sobre campos restritos do conhecimento. Para a solução de tais problemas, os SE precisam acessar uma substancial base de conhecimento (BC) do domínio da aplicação, que precisa ser criada do modo mais eficiente possível. Os SE devem, então, caracterizar-se por um conhecimento amplo e poderoso, organizado com o objetivo de simplificar a busca da resposta requerida.

Eles podem ser caracterizados como sistemas que reproduzem o conhecimento de um especialista adquirido ao longo dos anos de trabalho. Solucionam problemas que são resolvíveis apenas por pessoas especialistas (que acumularam conhecimento) na resolução destes problemas. Também são programas de computador que tentam resolver situações que os seres humanos resolveriam emulando o raciocínio de um especialista, aplicando conhecimentos específicos e inferências.

O conhecimento de um SE é organizado de tal forma que separa o conhecimento do domínio do problema e o conhecimento geral que abarca como resolver o problema. O conhecimento deve estar preparado para uma boa interpretação, e os objetos devem estar em uma determinada ordem representada por uma árvore de contexto. Possuindo o domínio do conhecimento separado, torna-se fácil para o analista desenhar procedimentos para a manipulação do conhecimento.

3.1 Especialista e engenheiro do conhecimento

O especialista é a pessoa que se consagra com particular interesse e cuidado a certo estudo ou ramo de sua profissão. Devido a seu conhecimento e experiência em determinada área, consegue realizar de forma eficiente, exata e precisa determinada tarefa. Ele possui um grande número de informações sobre determinada coisa e associada a ela, de forma direta ou não, permite abordar as causas do resultado de um determinado problema, como também tratar este problema de forma eficiente (RUSSELL & NORVIG, 2004).

O engenheiro do conhecimento procura investigar os SBC e suas aplicações, englobando atividades como: investigação teórica de modelos de representação do conhecimento, estabelecimento de métodos de comparação tanto do ponto de vista formal como experimental entre os diferentes modelos, desenvolvimento de SBC e estudo das relações entre sistemas e o processo ensino/aprendizagem (SAGHEB, 2006). Uma das tarefas mais difíceis do engenheiro do conhecimento é exatamente captar do especialista humano a estrutura do domínio do conhecimento. Dessa forma, o engenheiro do conhecimento deve ter uma visão clara do universo de conhecimento que ele irá extrair do especialista.

3.2 Diferenças entre SE e sistema convencional (SC)

A diferença entre um SC e um SE reside no fato de que o primeiro é baseado em um algoritmo, processa um conjunto de dados e instruções de forma repetitiva para emitir determinados resultados ao passo que um SE trabalha com heurística ao invés de algoritmo como também processa dados utilizando processos de inferência.

Os SE possuem facilidades em relação aos SC:

- Possibilidade para construção de regras.
- Tomada lógica de decisões sob imprecisão ou na ausência de informações.
- Nas aplicações (programas) tradicionais, o método de busca é baseado no conhecimento e nas regras codificadas previamente, havendo a necessidade de reescrita do código no caso do surgimento de novos conhecimentos. Já os SE podem recuperar novos fatos e regras e usá-los sem modificar a estratégia de busca.

3.3 Sistemas baseados em conhecimento (SBC)

Nos sistemas de informações tradicionais, o que se observa é uma eterna e penosa procura pelo que se deseja em meio a uma grande quantidade de informações emaranhadas. Sistemas de filtragem de dados esforçam-se para tornar estas tarefas mais amenas na tentativa de busca pelas informações de forma a subsidiar o usuário com as informações requeridas, a tempo e hora, para a tomada de decisão. É neste ponto que destacamos a eficiência dos SBC no gerenciamento da informação. Eles são capazes de receber informações de diversas origens e tipos, interpretá-las e analisá-las, identificando a sua pertinência e relevância, e direcioná-las para os diversos usuários de acordo com o interesse e a necessidade de cada um.

Os SBC são programas de computador que usam o conhecimento representado explicitamente para resolver problemas (FELFERING, 2006). Eles manipulam conhecimento e informação de forma inteligente e são desenvolvidos para serem usados em problemas que requerem uma quantidade considerável de conhecimento humano e de especialização. Assim, conhecimento e processo de resolução de problemas são pontos centrais no desenvolvimento de um SBC.

Em resumo, trata-se de um processo de análise de informação que procura reduzir o espaço de busca recuperando apenas as informações que são úteis para a resolução de problemas específicos. Para que o problema seja resolvido, o sistema deverá analisá-lo à luz das heurísticas armazenadas em seu motor de inferência e base de conhecimento e interagir com o usuário para obter todos os elementos, informações necessárias para a montagem do problema e possibilitar a busca de conhecimento para sua resolução.

Também é importante diferenciar os SBC dos SE. De uma forma geral, pode-se dizer que os SBC são sistemas capazes de resolver problemas usando conhecimento específico sobre o domínio da aplicação, ao passo que os SE são SBC que podem ser resolvidos por um especialista humano (WATERMAN, 1986). Por isso, eles requerem conhecimento sobre a habilidade, a experiência e as heurísticas usadas pelos especialistas. Na figura 3.1, são sintetizadas as principais características desses sistemas:

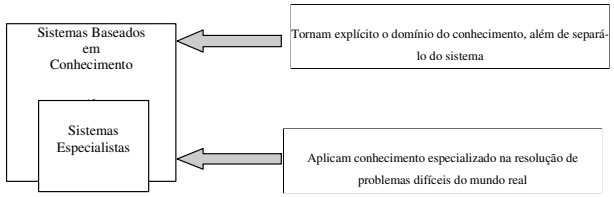


Figura 3.1 – SE e SBC (fonte: modificado Watermam, 1986).

Logo, SBC podem ser classificados como SE quando o desenvolvimento do mesmo é voltado para aplicações nas quais o conhecimento a ser manipulado restringe-se a um domínio específico e contam com um alto grau de especialização. Esses SE, construídos, principalmente, com regras que reproduzem o conhecimento do especialista, são utilizados para solucionar determinados problemas em domínios específicos.

Os SE começaram há 30 anos e se tornaram nos dias atuais realidade, sob a forma de sistemas interativos que respondem questões, solicitam e fornecem esclarecimentos, fazem recomendações, e geralmente auxiliam o usuário orientando-o no processo de tomada de decisão, ou seja, simulam o raciocínio humano fazendo inferências, julgamentos e projetando resultados. Assim, usuários e sistema caminham juntos, perguntando e fornecendo informações um ao outro até à completa solução do problema analisado.

3.4 Estrutura de um SE

WATERMAN (1986) sugeriu que o SE deveria conter a descrição do sistema sob duas perspectivas distintas: a do conhecimento processável pelo homem e a simbólica processável pelo computador.

Um SE apresenta em geral uma arquitetura com dois módulos, conforme mostrado na figura 3.2:

- Base de Conhecimento;
- Motor de Inferência.

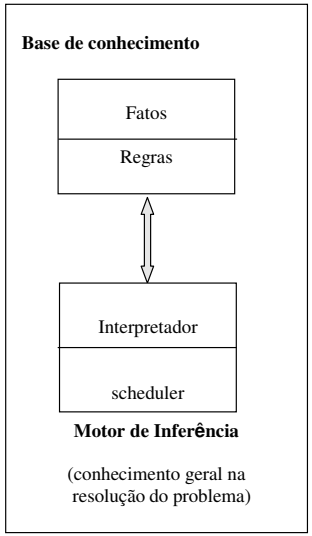


Figura 3.2 – Estrutura de um Sistema Especialista (WATERMAN, 1986).

Além de considerar a base de conhecimento e motor de inferência, os autores MOYNIHAM (2006) e HINGORANEY (1994) incluem o usuário como componente importante na estrutura de um SE.

3.4.1 Base de conhecimento (BC)

A BC é um elemento fixo, mas específico de um SE. É onde estão armazenadas as informações de um SE, ou seja, os fatos e as regras. Essas bases são implementadas pelo engenheiro do conhecimento, cujo papel é o de "extrair" procedimentos e estratégias de um especialista humano para a solução de determinado problema.

A BC desempenha papel essencial em qualquer sistema que se utiliza de agentes baseados em conhecimento. Tal base é representada por um conjunto de sentenças que, conforme salientam RUSSELL e NORVIG (2004), não devem ser confundidas com

sentenças gramaticais, pois são, na verdade, expressões técnicas reproduzidas em linguagem de representação do conhecimento.

O especialista toma decisões sobre determinado assunto com base em fatos que encontra e nas hipóteses que formula, ou ainda buscando em sua memória um conhecimento prévio armazenado sobre esses fatos e hipóteses. E o faz de acordo com sua experiência, isto é, com seu conhecimento acumulado sobre o assunto e, com esses fatos e hipóteses, emite a decisão.

3.4.2 Motor de inferência (MI)

O MI é um elemento essencial para a existência de um SE. É o núcleo do sistema. É por intermédio dele que os fatos, regras e heurística que compõem a BC são aplicados no processo de resolução do problema. Então, o MI define como o conhecimento será manipulado, porque é a parte responsável pela busca das regras da BC para serem avaliadas, direcionando o processo de inferência.

Basicamente, o MI é dividido em duas tarefas que são: o interpretador, que decide como aplicar as regras para inferir novos resultados; e o planejador, que decide quando e em que ordens às regras devem ser aplicadas. O MI opera como um "supervisor", tomando decisões e julgamentos baseados em dados simbólicos contidos na BC. Uma vez iniciado o sistema, cabe ao MI buscar na BC fatos e regras que serão comparados com as informações fornecidas pelos usuários. As regras definem relações lógicas entre conceitos no domínio do problema.

A decisão a ser tomada quanto ao motor de inferência também é conseqüência da aquisição de conhecimento. Sendo assim, é necessário verificar como o MI deve manipular o conhecimento, ou seja, como aplicar as regras aos fatos de maneira que represente fielmente o raciocínio do especialista.

3.5 Representação do conhecimento (RC)

Com base na descrição anterior, uma BC é um conjunto de representações de ações e acontecimentos do mundo. Cada representação individual é chamada de sentença. As sentenças são expressas em uma linguagem específica, chamada linguagem de RC.

A RC substitui o objeto ou fenômeno real, de modo a permitir a uma entidade determinar as conseqüências de um ato pelo pensamento, em vez de sua realização. Uma RC pode ser entendida como uma forma sistemática de estruturar e codificar o que se sabe sobre uma determinada aplicação. Uma RC deve apresentar as seguintes características:

- Ser compreensível.
- Ser robusta, isto é, permitir sua utilização, mesmo que não aborde todas as situações possíveis.
- Ser generalizável, ao contrário do conhecimento em si, que é individual. A RC é um dos problemas de IA, pois não existe uma teoria geral de RC, entretanto muitas técnicas de Representação do Conhecimento têm sido estudadas. A seguir são apresentadas brevemente algumas técnicas de RC.

3.5.1 Métodos baseados em regras

O Motor de Inferência processa a linguagem de representação usada na BC, gerando e percorrendo o espaço de busca sempre que necessário. Existem algumas linhas de raciocínio que podem ser seguidas pelos SBC. Por exemplo, no caso de regras de produção, existem:

- Encadeamento regressivo ou *backward chaining*: esse processo parte da suposição de que cada provável solução é verdadeira. Feito isso, tenta-se reunir evidências que comprovem ser correta a solução previamente considerada. Tais evidências são procuradas nas informações fornecidas pelo usuário.
- Encadeamento progressivo ou *forward chaining*: neste processo, as informações são fornecidas ao sistema pelo usuário, que, com suas respostas, estimulam o desencadeamento do processo de busca, explorando a BC, procurando pelos fatos, regras e heurísticas que melhor se aplicam a cada situação. O sistema continua nesta interação com o usuário até encontrar a solução para o problema a ele submetido.

A Representação do Conhecimento por intermédio desse método, é feita com pares de condição-ação. Se uma premissa IF (condição) é consistente para o problema, o sistema continua com a cláusula IF, tornando-a THEN (conclusão) para a próxima pesquisa na BC, até que encontre uma regra que o IF não seja considerada conclusão para outra regra. Ao mesmo tempo em que o sistema poderá iniciar uma nova pergunta ao usuário para obter informações adicionais

Segundo LANDAUER (1990) e MAK (2003), entre várias alternativas de RC, o método baseado em regras constitui uma forma natural de representar o conhecimento de um especialista humano.

3.5.2 Métodos baseados em redes semânticas e em *frames*

No formalismo de redes semânticas o conhecimento é representado utilizando-se uma estrutura de rede. Foi desenvolvida como um modelo cognitivo e tornou-se um método padrão de representação para IA e SE. Uma rede semântica consiste de nós, usualmente representando objetos (indivíduos, situações, conceitos em um domínio) conectados por ligações (arcos), representando as relações entre eles. Uma característica-chave da rede semântica é que importantes associações podem ser feitas explicitamente ou sucintamente, usando taxonomias (classe-de, faz-parte) bem estabelecidas para simplificar a resolução do problema. Outro esquema de representação de conhecimento desenvolvido na área de IA é chamado *frame* (estante). Em IA, o termo estante refere-se a um modo especial de designar um agrupamento de conhecimentos relevantes a objetos (indivíduo, alguma situação ou um conceito). Uma estante é organizada de maneira muito parecida com uma rede semântica, o conceito de nó é definido por uma coleção de atributos e valores destes atributos, onde os atributos são chamados de *slots* (prateleira) e cada prateleira possui por sua vez, um nome e consiste de um conjunto de atributos chamados facetas. Cada prateleira tem um número qualquer de procedimentos anexados a si, que são executados automaticamente quando a informação contida na prateleira é recuperada ou alterada.

Uma das principais características desse modelo de representação é a Herança de Propriedades, na qual uma classe mais especializada pode herdar todas as propriedades da classe mais geral. As associações entre estantes determinam a sua estrutura

hierárquica. Cada associação liga uma estante-pai ao seu filho. A estante-filho pode ser entendida como uma especialização da estante-pai.

Os SE são a solução de meu problema?

Uma questão importante que surge quando se discute sobre os SE se refere ao questionamento se os "SE são a solução de meu problema".

A resposta baseia-se na verificação de três aspectos:

- Desenvolvimento do SE: um SE tem seu desenvolvimento condicionado a fatores, tais como: a existência de especialista(s) possuindo uma abrangente experiência sobre determinada área de aplicação, também é necessário que os especialistas estejam de acordo entre si e a tarefa não deve transcender a complexidade cabível a um SE.
- Justificativa do desenvolvimento do SE: as características de um problema que justificam o desenvolvimento de um SE se referem, entre outras, a uma boa relação custo/ beneficio, à possibilidade de perda de conhecimento especializado (conhecimento tácito) e à necessidade de especialistas em zonas geográficas de difícil acesso.
- Adequação da tarefa: no sentido de examinar a natureza, a complexidade e o escopo do problema a ser resolvido.

Depois de definida a utilização de um SE para determinado problema, ainda discute-se sobre a manutenção de um especialista humano como parte do processo que envolve a utilização do sistema. A tabela 3.1. apresenta as vantagens e as desvantagens de ambos.

Como visto anteriormente, a elaboração de um SE envolve várias etapas, tais como: representação do conhecimento, motor de inferência, interface com usuário, aprendizagem e justificativa. Além disso, precisa-se da etapa de aquisição de conhecimento. Nem todos os sistemas baseados em conhecimento incluem todos estes itens, entretanto estes elementos constituem um sistema ideal para desempenhar uma ordem, já que seus frutos são interdependentes entre si.

Tabela 3.1 – Características dos especialistas humanos e dos SE (fonte: HART, 1986).

Especialista Humano	***Sistema Especialista***
Perecível	Permanente
Difícil de transferir Difícil de documentar	Fácil de ser transferido Fácil de documentar
Imprevisível	Consistente
Caro	Viável economicamente
Criativo	Sem inspiração
Adaptável	Deve ser atualizado
Sensorial	Alimentado com dados simbólicos
Visão ampla	Visão estreita
Bom senso	Conhecimento técnico

3.6 Aquisição do conhecimento

Alguns autores abordam o tema de aquisição do conhecimento como um processo que se divide em três etapas: decisão de qual é o conhecimento necessário; aquisição do conhecimento propriamente dito nos termos de extração do conhecimento do especialista e a representação do conhecimento extraído.

A tarefa de aquisição do conhecimento refere-se à transferência de conhecimento de alguma fonte, freqüentemente humana, para um programa de computador, isto é, de conhecimento tácito a conhecimento explícito. No contexto da construção de SE, a aquisição de conhecimento é o processo de captar conhecimentos, regras, métodos, enfim, o raciocínio do especialista de forma a entender e reproduzir a forma como ele resolve o problema para posteriormente transferi-lo para o sistema.

É fundamental que se compreenda o processo de raciocínio do especialista como um todo, para somente depois projetar a BC e aprofundar o nível de abstração.

A aquisição do conhecimento consiste de ações para reunir informações de um

ou mais especialistas humanos e/ou de fontes documentais, ordenando esta informação de alguma maneira e, então, traduzi-la para uma forma entendível pela máquina, ou seja, é o processo de transformar dados de especialistas em formalismo de implementação.

3.6.1 Método de aquisição do conhecimento

Devido às características dos métodos utilizados para AC, eles podem ser classificados entre quatro tipos básicos: intermediário, semi-intermediário, semi- direto, e direto supervisionado:

⇒ **Intermediário**: neste método, o engenheiro do conhecimento é o intermediário entre o conhecimento do especialista e a BC. O engenheiro do conhecimento atua de forma integral para a formação da BC. Este método é realizado por meio de entrevistas com o(s) especialista(s), estudo do problema e/ou pesquisas. Com base no conhecimento adquirido, o engenheiro do conhecimento codifica este conhecimento para a BC do sistema.

⇒ **Semi-intermediário**: neste método, o engenheiro do conhecimento é auxiliado por ferramentas computacionais para a aquisição de conhecimento de forma a auxiliá-lo neste processo. Tais ferramentas permitem ao engenheiro do conhecimento executar os procedimentos necessários de forma mais eficiente e/ou efetiva.

⇒ **Semi-direto**: aqui parte do trabalho do engenheiro de conhecimento é agora realizada de forma automática por intermédio de ferramentas utilizadas pelo(s) especialista(s), sendo que estas ferramentas interagem com o especialista para aquisição de conhecimento para a base. Tais ferramentas requerem treinamento dos especialistas, não somente para a sua utilização, mas também no processo de conhecimento.

⇒ **Direto Supervisionado**: a AC é realizada de forma automatizada através

de ferramentas que interagem com o especialista, sem a necessidade de participação do engenheiro do conhecimento ajudar na codificação de conhecimento para a base. No entanto, tal método precisa ser supervisionado pelo engenheiro do conhecimento para a validação do conhecimento adquirido.

3.7 Mecanismo de justificativa do SE

O mecanismo de justificativa é um requisito obrigatório nos SE, tendo, geralmente, capacidade de responder às seguintes perguntas:

⇒ Como chegou a essa conclusão?
⇒ Por que chegou a essa conclusão?
⇒ Por que não chegou à outra conclusão?

Os mecanismos de justificativa são capazes de descrever a linha de raciocínio empregada no sistema, o conhecimento que explica como o sistema chegou a suas conclusões e justifica os passos utilizados no processo. Alguns dos objetivos dos mecanismos de justificativa são: ensinar o usuário sobre o assunto, mostrar que sua conclusão é consistente e lembrar o usuário elementos importantes da análise que levam o sistema à determinada conclusão.

Este tipo de mecanismo torna o sistema mais confiável aos usuários e ainda representa um mecanismo de simulação, pois, tendo em vista uma alteração nos dados de entrada, pode-se verificar as conseqüências desta alteração no desenvolvimento do raciocínio.

3.8 Vantagens da utilização de SE

Dentre outras vantagens, podemos destacar:

⇒ O conhecimento dos especialistas pode ser distribuído, de forma que possa ser utilizado por um grande número de pessoas;

⇒ Um SE pode melhorar a produtividade e desempenho de seus usuários, considerando que o provê com um vasto conhecimento, que certamente, em

condições normais, demandaria mais tempo para assimilá-lo e, conseqüentemente, utilizá-lo em suas tomadas de decisão;

⇒ SE reduzem o grau de dependência que as organizações mantêm quando se vêem em situações críticas, inevitáveis, como, por exemplo, a falta de um especialista devido à mudança de fatores externos como: doença, morte, férias, entre outros. Ao registrar o conhecimento empregado nos SE, promove-se uma significativa redução no grau de dependência entre empresa e presença física do empregado;

⇒ SE são ferramentas adequadas para serem utilizadas em treinamentos de grupos de pessoas, de forma rápida e agradável, podendo servir, após o treinamento, como instrumento para coleta de informações sobre o desempenho dos treinados, obtendo subsídios para reformulação das lições e obtenção de melhor desempenho.

Conclusão

Os SE são um ramo da IA que buscam emular em um computador o raciocínio de um especialista de uma determinada área, bem como armazenar em uma BC todo o conhecimento relacionado a um problema específico.

O objetivo é a construção de sistemas de apoio à decisão chamado SE. Esse trabalho mostrou a construção deste tipo de ferramenta, inclusive apontando as diferenças e dificuldades inerentes às alternativas possíveis.

Os SE podem ser caracterizados como sendo programas computacionais que modelam a capacidade humana de resolução de problemas em domínios específicos do conhecimento, por meio de inferência lógica sob fatos e regras.

Os SE fornecem respostas a questões de uma área muito específica, fazendo inferências sobre conhecimento. Eles devem ser capazes de explicar a um usuário o seu processo de raciocínio e conclusões. Por isso, os SE podem fornecer "apoio à decisão" aos usuários na forma de um consultor especialista.

4 Resumo de um Processo Judicial Trabalhista

4.1 Introdução

Um processo trabalhista é a forma que um funcionário ou ex-funcionário tem de requerer junto ao judiciário especializado a reparação de uma suposta injustiça na aplicação da Consolidação das Leis do Trabalho (CLT), segundo a interpretação desta pessoa. Este capítulo pretende apresentar como este processo se origina, suas fases intermediárias, até o momento em que o funcionário tem seu pedido indeferido ou, ao contrário, recebe seu valor devido. Para uma melhor compreensão, este material será apresentado em ordem cronológica em relação aos fatos, peças jurídicas e instâncias do judiciário.

4.2 Origem – insatisfação do funcionário/ex-funcionário

A origem de um processo trabalhista, ou seja, Reclamação Trabalhista pode ocorrer de duas maneiras:

- A primeira quando um funcionário, chamado no jargão jurídico como reclamante, autor, ou pólo ativo, ainda se encontra exercendo sua atividade dentro da empresa, chamada comumente por reclamada, empresa ré, ou pólo passivo;
- E na segunda quando se trata de um ex-funcionário.

Em ambas as situações, o reclamante busca reparação financeira decorrente de alegadas perdas por parte da empresa ré ocorridas durante o contrato de trabalho, sendo sempre pedido uma indenização financeira, como, podendo ser requerido também uma reintegração ao emprego no caso de uma demissão indevida.

4.3 O Advogado

Para dar entrada em uma Reclamação Trabalhista, o reclamante tem a obrigação de constituir um advogado para representá-lo perante a Justiça. Logo, faz contato com

algum Advogado e descreve suas queixas. Este advogado, após entender a demanda do cliente, irá confeccionar e dará início à primeira peça do processo chamada de exordial.

4.4 Confecção da inicial (exordial)

Nesta fase, o advogado redige a peça que dará início ao processo, no qual consta a identificação do reclamante, último salário, e fundamenta todos os direitos que entender ser devido ao reclamante e não foram satisfeitos durante o contrato de trabalho. Neste último caso, quando se trata de um reconhecimento de vínculo empregatício. Ao final da exordial, o advogado descreve o chamado "rol de pedidos", que significa a sintetização de todos os pedidos elaborados durante a fundamentação. À exordial são incluídos alguns documentos, entre eles: procuração do reclamante para o advogado representá-lo; cópia da carteira de trabalho e outros documentos que comprovam o pedido em questão. O advogado também apresenta ou requer às provas que achar pertinente.

Ao final, o advogado protocola no Tribunal Regional do Trabalho (TRT) de sua comarca (região). Esta é uma das peças mais importantes no que tange a este trabalho de dissertação. O advogado descreveu e fundamentou a mesma, cada um dos pedidos e seus respectivos reflexos, que acha devido a favor de seu cliente e que serão julgados pelo Exmo. Juiz. Entende-se como pedidos de questões trabalhistas: horas extras, adicional de periculosidade, pedido de reintegração, pedido de vínculo trabalhista, equiparação salarial, entre outros. Como exemplo de reflexo, temos a integração das horas extras pedidas no fundo de garantia, férias e décimo terceiro salário. Elucidando um pouco mais a definição de reflexo, se o reclamante requer horas extras não pagas e se as mesmas forem deferidas, deverá ser pago também a correspondente integração das horas extras no fundo de garantia, férias e décimo terceiro salário, que talvez não tenham sido pagas porque o funcionário não tinha recebido as horas extras que pede neste momento. Sempre que é devida uma verba principal (hora extra, adicional de periculosidade, entre outros) serão devidos integrações/reflexos em verbas chamadas de acessórias (fundo de garantia, férias, 13º salário etc.). A figura 4.1 apresenta um "rol de pedidos" relativo a uma reclamação trabalhista.

RESUMO DE PEDIDOS NA INICIAL

A Reclamante, em sua peça inicial, pleiteia as seguintes verbas, *in verbis*:

Em face ao exposto, reclama, parcelas vencidas e vincendas:

- Pagto. das horas extras a serem apuradas;
- Pagto. do horário das refeições, acrescidos de 50% de acordo com o parágrafo 4º do artigo 71 da CLT;
- Diferença do adicional de insalubridade;
- Integração dos itens, "a", "b", e "c", nas férias vencidas 97/98, 98/99, nas gratificações natalinas, 1997, 1998, nas verbas rescisórias, FGTS, multa compensatória e no R.S.R;
- DAS VERBAS RESCISÓRIAS

e-1 Aviso prévio;

e-2 férias prop. 7/12 avos acrescidas de 1/3;

e-3 13º salário prop. 9/12 avos;

e-4 saldo de salário de 11 dias, em dobro, na forma do artigo 467 da CLT;

e-5 multa do artigo 477 da CLT;

e-6 FGTS sobre a rescisão;

e-7 TRCT código 01 e guia da CD;

e-8 40% sobre o FGTS;

e-9 honorários advocatícios a base de 20% sobre o valor da condenação;

e-10 Por último, baixa na CTPS, sob pena de multa equivalente a 1/30 da maior remuneração por dia de atraso.

Figura 4.1 – "Rol de Pedidos"

4.5 Da distribuição – ajuizamento

Ao receber a exordial, o setor de protocolo do TRT faz a distribuição (aleatoriamente, ressalvado os impedimentos devidos) deste processo a uma das Varas

do Trabalho. Ressalta-se que o Tribunal Regional do Trabalho é dividido por regiões, sendo o Rio de Janeiro sua primeira região (TRT 1ª Região), São Paulo, segunda região (TRT 2ª Região) etc. Cada TRT regional é subdividido em Varas do Trabalho – 1ª VT, 2ª VT etc.

A data protocolada da entrada do processo na Justiça é chamada de data de ajuizamento. Esta é uma data muito importante, não somente significa onde o processo ou a lide se inicia, mas também serve para a contagem dos juros que serão aplicados aos créditos do reclamante na fase de liquidação.

4.6 Notificação da reclamada

O próximo passo, já internamente na Vara Trabalhista, é notificar a reclamada para que a mesma tome ciência da reclamação trabalhista e apresente sua defesa.

4.7 Contestação

A contestação é a peça jurídica em que a empresa ré se defende das alegações do reclamante na exordial, sendo acompanhada de provas e documentos que julgarem relevantes e usa do direito à ampla defesa garantida pela Constituição do país.

4.8 Audiência

A primeira audiência na Vara do Trabalho normalmente é uma tentativa de conciliação, em se tratando de assuntos trabalhistas. Essa conciliação é feita de forma em que as partes envolvidas na ação tentem um acordo sobre a presente controvérsia. Se houver acordo, o processo tem seu fim naquele momento, do contrário, o Juiz abre prazo para que as partes apresentem novas provas e marca nova audiência quando serão ouvidas as testemunhas que, tanto reclamante como reclamada, indicaram nos autos da ação. Nesta audiência, também pode ser requerida pelas partes ou pelo Juiz a perícia de instrução para elucidar questões técnicas. Poderão ocorrer uma, duas ou até mesmo várias audiências, até que o Exmo. Juiz possa concluir a fase de conhecimento do processo.

A perícia é feita pelo perito judicial e tem papel de extrema relevância no processo trabalhista, pois é dele a responsabilidade de levantar dados técnicos para que

o Juiz tome a decisão sobre um ponto incontroverso, exemplo: se o reclamante requer equiparação salarial, o perito contábil deverá avaliar detalhadamente a função dos cotejados e instruir, por meio do laudo pericial, os advogados e o Juiz sobre aquilo que for o motivo da discordância.

Há outros tipos de perícia, como grafotécnica, periculosidade, insalubridades, médica, entre outras, que podem ser requeridas se forem necessárias para elucidar algum desacordo. A prerrogativa da perícia limita-se somente ao escopo técnico, isto é, o perito trabalha dentro de seu campo de atuação, limitando-se apenas a responder aquilo que for de sua natureza, sendo exclusivo do Juiz a decisão de concordar ou não com os fatos narrados pelo *expert*.

4.9 Sentença

Depois de realizada a última audiência e após o Juiz ler a exordial, a contestação, analisar as provas produzidas e ouvir o depoimento das testemunhas, do reclamante, bem como o depoimento do preposto da reclamada (pessoa que representa a empresa em Audiência), o Exmo. Juiz dará sua decisão, que poderá ser total ou parcialmente procedente ou improcedente o pedido do reclamante, fundamentado nas peças contidas nos autos da reclamação trabalhista. Ressalta-se que, a partir deste momento não é mais permitida a juntada de novas provas, seja de documentos ou testemunhas.

A sentença é composta das seguintes partes:

- ⇒ Abertura – data, nome das partes e seus respectivos representantes jurídicos (advogados);
- ⇒ Relatório – um breve resumo histórico dos fatos;
- ⇒ Fundamentação/Mérito – nesta fase, o Juiz fundamenta cada um dos pedidos requeridos pelo reclamante e, ao término, dá a decisão;
- ⇒ Dispositivo – fase final da sentença, o Juiz faz um resumo de tudo que transcreveu e decidiu na fundamentação.

A sentença é a peça jurídica mais importante na ligação do processo judicial com a proposta desta dissertação. Na seção chamada de fundamentação, o Exmo. Juiz irá

descrever, segundo o seu entendimento jurídico, o motivo pelo qual defere ou indefere cada um dos pedidos e reflexos solicitados pelo advogado da parte autora na peça exordial.

Este trabalho visa a utilizar as técnicas de Mineração de Textos (tarefa de classificação), Linguagem Convencional (LC) e Sistemas Especialistas (SE) para:

- Definir os pedidos que foram fundamentados na sentença;
- Identificar o Resultado de cada um destes pedidos (deferido/indeferido);
- Extrair cada uma das incidências (reflexos) geradas pelos pedidos;
- Capturar eventuais parâmetros para o cálculo de algum tipo de pedido;
- Utilizar as informações anteriores para que um SE calcule o valor devido ao cliente.

Este material trata do primeiro e último item descrito acima. Os outros itens serão desenvolvidos (implantados) em trabalhos futuros.

Com relação à definição dos pedidos (primeiro item acima), o mesmo tem por objetivo utilizar as técnicas de mineração de textos (MT) relacionadas às tarefas de classificação/categorização para identificar quais os pedidos (hora extra, adicional de periculosidade, equiparação salarial, vale transporte, entre outros) estão definidos na fundamentação do Exmo. Juiz. Isto é possível devido ao fato de que cada fundamentação de pedido pode ser decomposta em uma "bolsa de palavras" extraída da sentença. Com base nos métodos de mineração de textos, em que um grupo de "bolsa de palavras" passa por um algoritmo de aprendizado, usando técnicas de classificação, como SVM (Vapnik, 1999), Naive Bayes (McCallum, 1998), Rocchio (Rocchio, 1971), novas "bolsas de palavras" poderão ser classificadas. Os pedidos seriam as "classes" e estariam relacionados às "bolsas de palavras".

O segundo item se refere à identificação se este pedido foi deferido ou indeferido pelo Exmo. Juiz. Apesar de poder ser utilizada a MT para identificar este atributo do pedido, o melhor talvez fosse utilizar LC acoplada a um dicionário de dados (*thesaurus*), visto que outras palavras (procede, improcede, dou seguimento, não dou seguimento, entre outros) podem estar inclusas no texto em vez de defere/indefere.

Caso o pedido tenha sido deferido, o próximo passo seria capturar os reflexos que foram deferidos pelo Juiz. Entendem-se como reflexos, verbas (acessórios) que

devem ser calculadas sempre que existir um determinado tipo de pedido. Por exemplo, quando é deferida hora extra, esta acarreta também cálculo de reflexos, tais como: repouso semanal remunerado, fundo de garantia, décimo terceiro salário etc.; se são deferidos salários não pagos, poderia ter sido dado reflexo no fundo de garantia e décimo terceiro salário, mas não no repouso semanal remunerado. Esta operação poderia ser feita utilizando MT ou LC.

Posteriormente, a aplicação terá de capturar outros eventuais parâmetros que poderão estar junto da fundamentação daquele pedido como, por exemplo, o horário de trabalho do reclamante, caso o mesmo tenha tido horas extras como deferimento, o percentual de adicional de periculosidade, a data de início dos cálculos, entre outros. Esta última fase deve ser feita utilizando LC.

Ao final, depois de capturadas todas as informações básicas, o SE, com base em regras obtidas com os especialistas, irá calcular diversos valores, tais como: o valor que a reclamada deve ao reclamante; o valor que deverá ser pago à Previdência Pública (INSS) e o recolhimento a Receita Federal (IRRF). O cálculo destas verbas é o objetivo final deste trabalho.

Para uma melhor visualização, a figura 4.2 apresenta um trecho de uma sentença trabalhista contemplando a fundamentação do Juiz para cada pedido. Ressalta-se que esta sentença refere-se à fundamentação de um juiz específico. Cada juiz irá fundamentar a decisão de um pedido com um tipo de texto diferente. Podem ocorrer pequenas variações no formato estrutural entre os Juízes, mas, com certeza, na parte interna, irão fundamentar um pedido como "HORAS EXTRAS", com textos bastante diferentes. O próprio juiz do exemplo da figura 4.2, em outro processo que tenha também pedido de horas extras não irá repetir o mesmo texto. Essa grande diversidade de textos (dados não estruturados) associada ao imenso volume de documentos existentes, proporciona grande potencial para extrair conhecimento do texto tendo em vista a utilização das técnicas de mineração de textos.

4.10 Embargos

Os Embargos de Declaração são um recurso adicional à sentença de primeira instância, no qual as partes podem pedir um esclarecimento de uma decisão obscura, não clara, ou quando o juiz deixa de apreciar algum pedido ou alguma contestação

(impugnação) feita no decorrer do processo. Finda-se neste momento o que é chamado de 1ª instância, a decisão do Juiz por meio da Sentença e dos Embargos.

SENTENÇA

(...)

HORAS EXTRAS

Afirma a Reclamante que desenvolvia trabalho, de segundas a sextas-feiras, no horário de 8h às 18h, do início do contrato a julho/1998, passando depois para prestação de serviços em dias alternados das 7h às 19h30 min, sempre sem intervalo para refeição, não recebendo pagamento por serviços extraordinários.

Defende-se o Reclamado informando inexistência de horas extras, afirmando jornada das 8h às 18h de segundas a quintas-feiras e das 8h às 17h nas sextas-feiras, com posterior alteração para escala de 12x36, das 7h às 19h, sempre com 1 hora de intervalo.

O Reclamado junta controles de horário, fls. 23/25, onde fica comprovado o horário alegado na defesa, quanto ao início e término de jornada, não havendo registro de intervalos, alegando que estes não precisam ser registrados, com invocação de norma administrativa indicada em defesa.

O horário de trabalho deve ser registrado, inclusive quanto aos intervalos, para fins de comprovação em juízo pelo empregador. Em não sendo acolhe-se o afirmado na exordial, quanto à inexistência de intervalos.

Registre-se que o Reclamado oferece defesa, no que respeita à jornada de trabalho, não invocando o instituto da compensação, sem comprovação de existência de contrato neste sentido.

A legislação estabelece como limite diário de trabalho 8h, sendo extras todas as horas trabalhadas em horário superior, com adicional de 50%, inexistindo previsão legal para a jornada de trabalho praticada pela Autora.

Assim, conforme prova nos autos, acolhe-se o horário indicado em defesa, como sendo de 8h às 18h, de segundas a quintas-férias e das 8h às 17h, nas sextas feiras, até julho /1998, passando após, até o final do contrato, para 7h às 19h, em escala de 12x36, **condenando-se a Reclamada no pagamento de horas extras, com adicional de 50%, sobre o trabalho prestado após a 8ª hora diária, de segundas a sextas-feiras, com integração, por habituais, à remuneração de repousos semanais, 13º salários, férias com adicional de 1/3, aviso prévio e multa do art. 477, § 8º, da CLT.**

Pela ausência de comprovação de intervalo de descanso e refeição de 1 hora, **defere-se o adicional de hora extra de 50% ao dia, incidente sobre 1 hora de salário, em todos os dias de trabalho ao longo do contrato. É devido apenas o adicional, porque a hora normal já está paga, sendo utilizado o mesmo entendimento do Enunciado n. 85, do Colendo TST.**

FGTS E SEGURO-DESEMPREGO

O Reclamado é **condenado em indenização de FGTS, inclusive multa de 40%, incidente sobre as horas extras e adicional de horas extras deferidos, bem como sobre aviso prévio e 13º salário proporcional de 1999.**

Determina-se a imediata entrega da guia para saque de FGTS, sob pena de multa diária no valor equivalente a 1/10 do último salário da Autora, por força do art. 461, §4º, do Código de Processo Civil, aplicado de forma subsidiária.

Estabelece a lei que compete ao empregador fornecer a seu empregado às guias próprias para a percepção do benefício do seguro desemprego, quando da despedida. Em virtude do descumprimento de tal obrigação na época própria, **condena-se o Reclamado no pagamento de indenização equivalente ao direito, com base no art. 159 do Código Civil, aplicado de forma subsidiária, conforme pedido.**

As alegações de defesa são impertinentes, sendo devida a indenização, eis que o empregador deu causa ao não recebimento, por descumprimento de obrigação de fazer.

COMPENSAÇÃO

Inexistem compensações a serem deferidas, porque o Reclamado não comprovou pagamentos dos valores objetos da condenação.

PRESCRIÇÃO

Inexiste prescrição a ser declarada.

ASSISTÊNCIA JUDICIÁRIA GRATUITA e

HONORÁRIOS DE ADVOGADO.

A Reclamante percebia salário mensal inferior ao dobro do salário mínimo legal, fazendo jus ao benefício de Assistência Judiciária Gratuita, deferido de ofício, diante da previsão legal aplicável, contida na Lei n. 5.584/70, combinada com a Lei n. 1.060/50.

Considerando a Assistência Judiciária Deferida o Advogado atuou em serviço à Justiça e ao Estado, não sendo exclusivo do Sindicato Profissional o exercício de tal encargo, porque a parte tem o direito de escolher o profissional de sua preferência.

Na forma da Lei, 1.060, de 5/2/1950, art. 11, **condena-se o Demandado no pagamento de honorários de advogado de 15% sobre o valor total da condenação, a ser apurado em liquidação de sentença.**

.....

Figura 4.2 – Exemplo de uma sentença.

4.11 Recurso Ordinário

Inicia-se neste momento o que é chamado de 2ª instância. É com base no Recurso Ordinário que as partes têm a oportunidade de reformar a sentença, ou seja, mudar o julgado na 1ª instância. Assim como na distribuição do processo para a vara do trabalho que tratará desta ação judicial, no recurso também há sorteio de uma turma que irá julgar os recursos interpostos pelas partes. A composição do recurso ordinário deve ser feita nos mesmos moldes da sentença, apenas com um diferencial, antes de ir a julgamento passa por um relator, se este não aprovar, não haverá julgamento do mérito, aprovando os autos, vão a julgamento por uma turma de Juízes conhecidos como Desembargadores (são compostos de três ao todo).

Após análise dos autos, os mesmos votam e o relator transcreve a decisão da maioria em uma peça chamada de acórdão. O acórdão também é uma peça muito importante para a aplicação deste trabalho, ela também pode modificar qualquer dos parâmetros minerados anteriormente na sentença do juiz de 1ª instância. Assim, deve passar pelas mesmas técnicas de mineração de textos (classificação) processados para a sentença, ou seja, primeiramente serão identificados os tipos de pedidos, a seguir, se foram deferidos ou indeferidos, no próximo passo, os reflexos e, por último, os outros parâmetros adicionais. O formato estrutural do acórdão é muito parecido com o da sentença, tal como mostrado na figura 4.2, mas ressalta-se que a fundamentação (texto) é sempre muito diferente.

4.12 Embargos do acórdão

Assim como na 1ª instância, o recurso ordinário também tem seus embargos de declaração, com a mesma finalidade de esclarecimento ou julgar aquilo que ficou omisso.

4.13 Recurso de Revista

Conhecido como 3ª instância. Neste momento, só se podem discutir pontos unicamente interpretativos ou alguma ofensa às leis em vigor. Neste ato, só reforma-se a

sentença se comprovado os dois casos acima, do contrário, permanecem as decisões anteriores.

O recurso de revista tem sua composição de forma igual a do recurso ordinário, com alguns diferenciais. Primeiro, o processo é analisado pelo vice ou presidente do TRT regional, onde se encontram os autos. Este desembargador é quem faz a primeira análise, se os autos deverão subir ou não à 3ª instância, ou seja, Tribunal Superior do Trabalho (TST), em Brasília. Em caso de positivo, o processo é encaminhado. Um novo relator faz uma análise, sendo os autos aprovados, vai-se a julgamento por outra junta, que é formada por sete desembargadores e que, após a votação, retorna ao relator para transcrição da decisão vencedora no voto.

4.14 Embargos

O procedimento de embargos de declaração nesta fase é idêntico aos Embargos do acórdão e da Sentença.

4.15 Agravo de petição

Recurso para fase de execução trabalhista, interposto em fase de decisão definitiva (sentença), com matéria delimitada, geralmente contra a decisão de embargos à execução ou de terceiro, no juízo *a quo* para ser examinado pelo juízo *ad quem*.

4.16 Artigos de liquidações

Após toda a tramitação processual em todas as instâncias, os autos retornam à vara do trabalho de origem para dar continuidade. Esta fase consiste em apresentar valores devidos sobre a coisa julgada, em outras palavras, sobre as verbas deferidas e confirmadas ou alteradas por meio dos vários instrumentos jurídicos já citados. Normalmente, o reclamante apresenta primeiro os cálculos que acha devido. O Juiz, ao recebê-los, notifica a reclamada para que a mesma avalie estes cálculos, concorde ou apresente novos cálculos se discordar. As partes tentarão chegar a um acordo. Caso não aconteça, o Juiz irá decidir ou poderá requerer perícia contábil para que um perito realize os cálculos. A perícia também poderá ser requerida caso exista complexidade

nos cálculos. Durante a realização da perícia, o *expert* poderá requerer documentos que não se encontram nos autos, bem como realizar diligências a empresa ré, ou a qualquer outro local que se fizer necessário para elucidação dos fatos. Cabe destacar que, na fase de liquidação, somente o perito poderá incluir novos documentos nos autos.

Após a entrega do laudo pericial, o Juiz abre novo prazo às partes para se manifestarem sobre o laudo. Conseqüentemente, o perito também terá prazo para responder se houver alguma impugnação das partes. Se a divergência persistir, o Juiz analisará os manifestos do perito e das partes e tomará a decisão sobre os cálculos que achar correto.

Este momento em que se calcula o valor devido ao reclamante é a segunda parte, que relaciona um processo trabalhista a esta dissertação. O vínculo decorre da necessidade de se fazer cálculos de forma repetitiva, objetivando rapidez, evitando erros e sendo necessário utilizar grande quantidade de regras. Isto nos remete para os Sistemas Especialistas (*Expert Systems*).

Ressalta-se que o objetivo completo deste trabalho é aplicar as técnicas de Mineração de Textos (MT) para classificar os pedidos feitos pelos reclamantes, com base nas fundamentações existentes nas sentenças proferidas pelos Juízes. A seguir, utilizar LC para identificar se cada pedido foi deferido ou indeferido. Se o pedido foi deferido, utilizar também LC ou MT para capturar os reflexos e outra vez LC para capturar outros parâmetros necessários ao cálculo. Depois, estas informações são passadas para um Sistema Especialista, via uma interface, que, com base em regras obtidas anteriormente com especialistas, irá calcular o valor exato que a reclamada deve ao autor, também o que deverá ser pago à Previdência Social (INSS) e à Receita Federal (IRRF).

4.17 Embargos à execução

Com a homologação dos cálculos, o Juiz abre novo prazo às partes para se manifestarem sobre o julgado. Entretanto, o Juiz notificará a parte devedora a efetuar o pagamento dos valores incontroversos, ou quitação total. Se quitado e não houver embargos do decisório, o processo é encaminhado para arquivo, caso contrário, continua-se a discutir, porém o valor depositado é imediatamente liberado à parte credora, dando-se, assim, prosseguimento nos autos.

Os embargos de execução irão até não haver mais o que se discutir, ou o Juiz entender que todas as peças jurídicas satisfazem sua decisão, encerrando o processo, com os recolhimentos tributáveis devidos e o pagamento ao reclamante.

5. Descrição do Sistema

O sistema completo é composto de várias etapas técnicas, envolvendo, desde Mineração de Textos (MT), passando por Linguagem Convencional até Sistemas Especialistas (SE). Contempla também trabalhos manuais, como: seleção de sentenças na Internet, utilização de arquivos digitados, entre outros. Este capítulo tem por objetivo apresentar a descrição de todas as etapas necessárias para a execução da aplicação utilizada neste trabalho.

5.1 Obtenção dos dados para escolha do melhor classificador

Os dados iniciais estavam originalmente em processos (papel) trabalhistas da 1ª Região (Rio de Janeiro). No trabalho do perito, é necessária a digitação do laudo pericial, que contém as peças trabalhistas (sentença, acórdão, embargos, entre outros) utilizadas no processamento desta aplicação. O laudo é digitado em arquivo *word*. Deste arquivo foram criadas "bolsas de palavras", tendo por base o texto que fundamenta cada tipo de pedido. Um exemplo de sentença foi apresentado na figura 4.2. A respectiva divisão da mesma em três "bolsas de palavras" (BP) pode ser vista nas figuras 5.1 a 5.3. Todo o texto inserido entre as palavras em maiúsculas será considerado uma BP, relativa ao pedido representado pelo primeiro tópico em maiúsculo.

Em que pese ter sido apresentado acima uma BP que é referente à FGTS e Seguro Desemprego, ressalta-se que neste trabalho foram geradas e utilizadas BP de somente quatro tipos de pedidos: alimentação, equiparação salarial, horas extras e honorários de advogado, apesar de existirem outros tipos de pedido, como: periculosidade, insalubridade, comissões, repouso semanal remunerado (RSR), vale transporte, verbas natalinas, verbas rescisórias, entre outros. Foram utilizados somente quatro tipo de pedidos em função de ser necessário escolher uma amostra para validação do trabalho.

(....)

HORAS EXTRAS

Afirma a Reclamante que desenvolvia trabalho, de segundas a sextas-feiras, no horário de 8h às 18h, do início do contrato a julho/1998, passando depois para prestação de serviços em dias alternados das 7h às 19h30 min, sempre sem intervalo para refeição, não recebendo pagamento por serviços extraordinários.

Defende-se o Reclamado informando inexistência de horas extras, afirmando jornada das 8h às 18h de segundas a quintas-feiras e das 8h às 17h nas sextas-feiras, com posterior alteração para escala de 12x36, das 7h às 19h, sempre com 1 hora de intervalo.

O Reclamado junta controles de horário, fls. 23/25, onde fica comprovado o horário alegado na defesa, quanto ao início e término de jornada, não havendo registro de intervalos, alegando que estes não precisam ser registrados, com invocação de norma administrativa indicada em defesa.

O horário de trabalho deve ser registrado, inclusive quanto aos intervalos, para fins de comprovação em juízo pelo empregador. Em não sendo acolhe-se o afirmado na exordial, quanto à inexistência de intervalos.

Registre-se que o Reclamado oferece defesa, no que respeita à jornada de trabalho, não invocando o instituto da compensação, sem comprovação de existência de contrato neste sentido.

A legislação estabelece como limite diário de trabalho 8h, sendo extras todas as horas trabalhadas em horário superior, com adicional de 50%, inexistindo previsão legal para a jornada de trabalho praticada pela Autora.

Assim, conforme prova nos autos, acolhe-se o horário indicado em defesa, como sendo de 8h às 18h, de segundas a quintas-férias e das 8h às 17h, nas sextas feiras, até julho /1998, passando após, até o final do contrato, para 7h às 19h, em escala de 12x36, **condenando-se a Reclamada no pagamento de horas extras, com adicional de 50%, sobre o trabalho prestado após a 8ª hora diária, de segundas a sextas-feiras, com integração, por habituais, à remuneração de repousos semanais, 13º salários, férias com adicional de 1/3, aviso prévio e multa do art. 477, § 8º, da CLT.**

Pela ausência de comprovação de intervalo de descanso e refeição de 1 hora, **defere-se o adicional de hora extra de 50% ao dia, incidente sobre 1 hora de salário, em todos os dias de trabalho ao longo do contrato. É devido apenas o adicional, porque a hora normal já está paga, sendo utilizado o mesmo entendimento do Enunciado n. 85, do Colendo TST.**

(....)

Figura 5.1 – "Bolsa de palavras" extraída da sentença relativa ao pedido "hora extra".

(....)

FGTS E SEGURO-DESEMPREGO

O Reclamado é **condenado em indenização de FGTS, inclusive multa de 40%, incidente sobre as horas extras e adicional de horas extras deferidos, bem como sobre aviso prévio e 13º salário proporcional de 1999.**

Determina-se a imediata entrega da guia para saque de FGTS, sob pena de multa diária no valor equivalente a 1/10 do último salário da Autora, por força do art. 461, §4º, do Código de Processo Civil, aplicado de forma subsidiária.

Estabelece a lei que compete ao empregador fornecer a seu empregado às guias próprias para a percepção do benefício do seguro desemprego, quando da despedida. Em virtude do descumprimento de tal obrigação na época própria, **condena-se o Reclamado no pagamento de indenização equivalente ao direito, com base no art. 159 do Código Civil, aplicado de forma subsidiária, conforme pedido.**

(....)

Figura 5.2 – "Bolsa de palavras" extraída da sentença relativa ao pedido FGTS e Seguro Desemprego.

(....)

ASSISTÊNCIA JUDICIÁRIA GRATUITA e HONORÁRIOS DE ADVOGADO

A Reclamante percebia salário mensal inferior ao dobro do salário mínimo legal, fazendo jus ao benefício de Assistência Judiciária Gratuita, deferido de ofício, diante da previsão legal aplicável, contida na Lei n. 5.584/70, combinada com a Lei n. 1.060/50.

Considerando a Assistência Judiciária Deferida o Advogado atuou em serviço à Justiça e ao Estado, não sendo exclusivo do Sindicato Profissional o exercício de tal encargo, porque a parte tem o direito de escolher o profissional de sua preferência.

Na forma da Lei, 1.060, de 5/2/1950, art. 11, **condena-se o Demandado no pagamento de honorários de advogado de 15% sobre o valor total da condenação, a ser apurado em liquidação de sentença.**

(....)

Figura 5.3 – "Bolsa de palavras" relativa ao pedido "honorários advocatícios".

Cada documento a ser processado é gerado a partir das peças jurídicas (sentenças ou acórdãos) e cada um deles pode conter várias "bolsas de palavras" (BP) de tipos de pedidos diferentes. No trabalho, como um todo, foram geradas 104 BPs, relativas aos quatro tipos de pedidos referenciados oriundos de várias peças jurídicas e proferidas por vários Juizes das 06, 07, 13, 55, 69 Vara Trabalhista (VT) do Estado do Rio de Janeiro e da 2ª instância. É importante frisar que, devido à necessidade de um grande número de BP, além dos documentos *scaneados* ou digitados, foi necessário complementar com sentenças e acórdãos capturados na Internet (http://www.7vtrj.com), relativos à 7ª VT RJ. A relação das BPs com esta dissertação é devido às mesmas serem os arquivos de treinamento e teste para os algoritmos de classificação de Mineração de Textos (MT), e os pedidos são as classes/categorias. Este grande volume BPs relativas a um mesmo pedido acarreta uma diversidade de agrupamento de palavras bastante positiva para a aplicação das técnicas de MT.

5.2 Preparação dos dados

O próximo passo foi separar os arquivos em treinamento e teste. A seguir, com base na execução de um programa Java que concatena os vários arquivos existentes, foi gerado um único arquivo de treinamento, e outro de teste, ambos com todas as classes (pedidos).

Empregou-se validação cruzada com subamostragem aleatória (*Random Subsampling*) da seguinte forma. Foram geradas cinco combinações aleatórias diferentes com as 104 BPs (arquivos) nos dois arquivos de treinamento/teste, e os resultados finais foram obtidos através de uma validação cruzada com subamostragem aleatória (*random subsampling*).

Posteriormente, estes dois arquivos foram convertidos para formato XML que os torna mais simples para manipulação em aplicações de MT. Um exemplo deste arquivo pode ser visto na figura 2.3.

Com os arquivos em formato XML, para retirada de *stopwords*. O anexo A apresenta uma lista de *stopwords*. Conforme detalhado no capítulo 2, a retirada de *stopwords* tem como finalidade reduzir a grande dimensionalidade das aplicações de MT que requerem grande espaço para armazenamento dos dados e alta capacidade de

CPU. Portanto, foram excluídas palavras desnecessárias do tipo artigos, preposição, conjunções, pronomes, tais como: de, assim, afim, agora, onde, outro, outros, ainda, a, o, que, entre outros.

O último passo para terminar a preparação dos arquivos para processamento foi à execução de um algoritmo de *stemming*. Este procedimento reduz a quantidade de palavras diferentes no texto por intermédio de uma lógica que leva em consideração as características de cada linguagem para retirar sufixos e gerar palavras com o menor radical possível. Dentre os vários algoritmos de *stemming* existentes, entre eles o Método de Lovins (LOVINS, 1968), Stemmer S, Método de Porter (PORTER, 1980) e RSLP (ORENGO, 2001), detalhados no capítulo 2, foi escolhido para ser utilizado neste trabalho o RSLP.

5.3 Processamento da parte referente à Mineração de Textos

Depois de geradas as cinco combinações diferentes de arquivos "treinamento/teste" com as bolsas de palavras, em formato XML, retiradas as *stopwords* e executada a rotina de *stemming*, o próximo objetivo consiste no processamento destas informações para geração do melhor classificador possível, utilizando os arquivos de treinamentos, validando contra os arquivos de teste, por meio das técnicas de MT. Este classificador será utilizado para identificar corretamente cada tipo de pedido de um novo documento composto de várias "bolsas de palavras". Na próxima seção, serão descritos os aplicativos utilizados neste trabalho.

5.3.1 Text-Miner Software Kit (TMSK)

O TMSK é uma ferramenta computacional para tarefas de MT, tais como: Classificação, Recuperação de Informações, Procurando Estruturas, Extração das Informações. Cada uma destas aplicações requer anteriormente uma fase de preparação dos dados. Existe também um ou mais serviços para estas tarefas e cada um destes serviços pode ser composto de uma ou mais rotina. A figura 5.4 apresenta a relação entre as tarefas, serviços e rotinas.

Preparação dos Dados	Serviços	Rotinas TMSK
	Criação do Dicionário	*mkdict*
	Criação do Vetor	*vectorize*

Tarefas de MT	Serviços	Rotinas TMSK
Predição/Classificação	Naive Bayes	*nbayes, testnbayes*
	Modelo Linear	*linear, testline*
Recuperação da Informação	Documento/Consulta que conferiu	*matcher*
Procurando Estruturas	*Clustering K-means*	*kmeans*
Extração de Informações	Identificação do nome da Entidade	*tagNames*

Figura 5.4 – Serviços e Rotinas do TMSK.

Como um dos objetivos do trabalho é a classificação de documentos, iremos nos limitar a descrever abaixo somente os serviços e rotinas referentes a este tipo de aplicação.

O primeiro passo para execução do TMSK é a geração de um dicionário de dados para cada um dos cinco arquivos (treinamento/teste) com as "bolsas de palavras". O formato deste arquivo pode ser visto na figura 5.5.

A seguir é gerado um arquivo de vetor do tipo "esparsos", baseado nas palavras selecionadas pelo dicionário dados e nos arquivos de entrada. A figura 5.6 apresenta um arquivo com este formato.

A próxima etapa consiste na construção de um classificador para cada uma das duas técnicas disponíveis no TMSK, Naive Bayes e Linear por Ordenação, detalhadas no capítulo 2.

Ao final é executada a rotina com o arquivo de teste que irá avaliar o desempenho dos classificadores e gerar dois arquivos, um com os documentos que são "classe" e o outro com as não "classe".

```
extras
reclamante
fls
intervalo
prova
adicional
jornada
hora
horário
fato
controles
cartões
natalina
50%
gratificação
art
função
regional
violação
freqüência
minutos
repouso
salário
remunerar
autor
período
```

Figura 5.5 – Dicionário de Dados gerado para a classe “horas extras”.

```
0 1@1 4@2 7@1 16@1 18@1 23@3 24@1 25@2 31@2 45@2 46@1 52@1 59@2
0 1@1 8@1 16@1 23@3 24@1 34@1 36@2 41@1 45@2 52@1 59@1 63@1 72@1
0 1@1 7@1 8@1 21@1 23@1 25@1 27@3 29@2 31@1 33@1 34@1 45@1 53@2
0 3@2 8@1 9@1 12@1 17@1 18@2 41@1
0 5@1 9@1 19@1 21@2 23@1 27@4 29@1 52@1 70@3 74@1 75@3 79@6 114@1
0 4@1 6@1 8@1 9@1 16@2 23@2 29@1 39@1 56@1 62@1 69@1 75@2 84@1
0 2@1 3@1 4@2 6@1 8@1 10@2 16@2 17@1 18@1 20@1 21@2 41@1 42@1
0 3@1 4@1 6@1 12@1 16@2 17@1 21@1 23@1 41@1 42@1 48@1 51@1 55@1
0 9@1 22@1 75@1 204@1 210@1
0 7@1 12@1 16@1 18@3 23@2 33@1 41@1 48@1 52@1 75@1 138@1 141@1
0 9@1 16@1 17@1 18@1 29@1 36@1 37@1
1 1@4 2@4 4@1 6@1 7@4 10@5 12@1 13@3 14@6 15@6 16@1 17@2 21@1
1 1@13 2@4 3@7 4@7 5@3 6@6 7@1 9@4 11@6 12@5 13@1 14@1 15@2 16@4
1 1@3 2@2 16@1 23@1 30@1 34@1 54@1 67@1 94@2 134@1 161@1 162@1
1 1@3 2@3 7@2 8@3 10@1 13@4 14@2 15@1
1 1@1 3@1 4@1 10@1 13@1 15@1 17@1 18@1 22@1 24@2 25@1 26@1 31@1
```

Figura 5.6 – Arquivo de vetores esparsos.

5.3.2 Rule Induction Kit for Text (RIKTEXT)

RIKTEXT é um pacote de *software* para indução de regras de decisão com o objetivo de classificar documentos. Em vez de números complexos como os gerados pelo classificador Linear e pelo *Naive Bayes*, este modelo apresenta regras de lógica simples e facilmente interpretáveis. Como exemplo, podemos citar que uma “bolsa de palavras” (BP) que contém a palavra “horas extraordinárias” deve ser classificada como pedido (classe) “horas extras”. Em contrapartida, uma BP que contém “alimentos” deve ser classificada com pedido “alimentação”. Portanto, este tipo de classificador tem por objetivo encontrar o melhor conjunto de regras utilizando as palavras existentes no texto para fazer classificações. O melhor conjunto de regras será aquela com a menor quantidade de regras e com o menor erro. A figura 5.7 apresenta as regras obtidas com o RIKTEXT para uma das execuções relativas ao pedido de “horas extras”.

```
Ruleset made using no prune mode. [0,0,5]
hext
~hext
extras & fls --> hext
horário --> hext
extraordinárias --> hext
[TRUE] --> ~hext
```

Figura 5.7 – Regras do aplicativo RIKTEXT para um pedido do tipo “hora extra”.

Tal como no aplicativo TMSK, o RIKTEXT também requer inicialmente a geração de um arquivo de dicionário de dados e um do tipo vetor, como também gera ao final dois arquivos separando as classes e não classes como também a performance do classificador.

5.3.3 Escolha do melhor classificador

Após processar os classificadores gerados pelo TMSK (método Linear e *Naive Bayes*) e pelo RIKTEXT (Indução de regras) com os arquivos de treinamento/teste citados, é escolhido o melhor classificador em função dos indicadores de desempenho definidos no capítulo 2 (precisão e cobertura).

5.4 Processamento da parte referente ao Sistema Especialista

Após serem utilizadas as técnicas de MT para identificar cada pedido e o programa de linguagem tradicional ter capturado outros parâmetros necessários, a última parte do sistema é composta de um SE, que tem como objetivo:

- Calcular o valor final devido pela reclamada ao reclamante;
- O valor que deverá ser pago à Previdência Pública (INSS);
- O recolhimento de Receita Federal (IRRF).

O cálculo destas verbas é o objetivo final deste trabalho. O desenvolvimento do sistema foi feito em HTML, ASP e Javascritpt com banco de dados ACESS e procura ser parametrizado ao máximo para facilitar o usuário. A seguir, serão descritos os vários elementos/módulos do sistema.

5.4.1 Tabelas

Vários repositórios específicos com características de "tabelas" de dados são utilizados pelo SE para efetuar os cálculos, como: Alíquotas de Imposto de Renda, Alíquotas do INSS, Histórico do Salário Mínimo, Valor da Correção Monetária, Valores da URV, Tabela de faixa do Seguro Desemprego, Tabela de Feriados nacionais e locais, entre outros.

Entende-se por tabelas arquivos auxiliares que não são gerados pelo sistema, mas sim consultados por este durante o processamento com o objetivo de se obter uma informação auxiliar. A grande maioria destas informações é inserida e mantida por meio da digitação no próprio sistema. Algumas tabelas, como as de Alíquotas de Imposto de Renda e Alíquotas do INSS, poderiam ser capturadas no site do respectivo órgão, mas, por motivos de "custo/beneficio", optou-se pela informação ser digitada. A tabela de Valor da Correção Monetária pelo fato de ser alterada todos os meses com valores relativos há aproximadamente trinta anos, é a única obtida a partir de um *download* do site do Conselho Nacional da Justiça do Trabalho (http://informatica.jt.gov.br/portal/page?_pageid=135,161405&_dad=portal&_schema=PORTAL). A figura 5.8 apresenta o menu de tabelas.

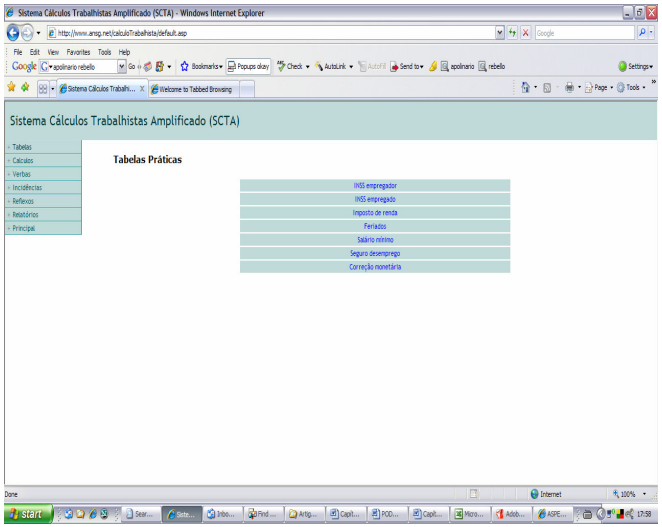


Figura 5.8 – Menu principal e o de tabelas do sistema.

5.4.2 Dados iniciais e externos ao processo

Além das informações obtidas junto à fundamentação das sentenças dos Juízes (pedidos por meio das técnicas de MT e outros parâmetros via LC) como também das tabelas anteriormente citadas é necessário inserir no SE alguns outros dados para seu processamento. Estes dados podem ser classificados de duas formas: o primeiro deles diz respeito a informações avulsas que podem ser definidas como atributos de cada processo. Exemplificando: nome do reclamante e da reclamada, data de admissão e demissão, data do ajuizamento, se tem contribuição a fundo de pensão, entre outros. O segundo tipo de informação utilizada quase sempre é necessário no processo, que se caracteriza por ter periodicidade mensal, é normalmente o salário e algumas outras verbas salariais, que juntas completam a chamada remuneração de um funcionário, tais como: anuênio, qüinqüênio, gratificação de função, abono salarial, ajuda de custo etc. Algumas vezes é utilizado algum outro tipo de informação que não seja do tipo salarial, entre eles os horários de entrada e saída de um cartão de ponto quando um Juiz defere hora extra com base no mesmo.

5.4.3 Processamento do SE

Com todas as informações obtidas até o momento e inseridas no SE, o mesmo, com base em regras obtidas com especialistas, irá fazer o processamento e gerar as saídas esperadas. Seguem abaixo os relatórios emitidos pelo SE.

.

- Demonstrativo das Horas Extras – mostra o quantitativo de horas extras totalizadas por mês, levando-se em conta os dias efetivamente trabalhados e os feriados. Os casos em que é necessária esta totalização acontecem quando existe cartão de ponto e os horários são diários. Esta saída somente é gerada se houver deferimento do Juiz deste tipo de pedido. A figura 5.9 mostra o demonstrativo de horas extras.

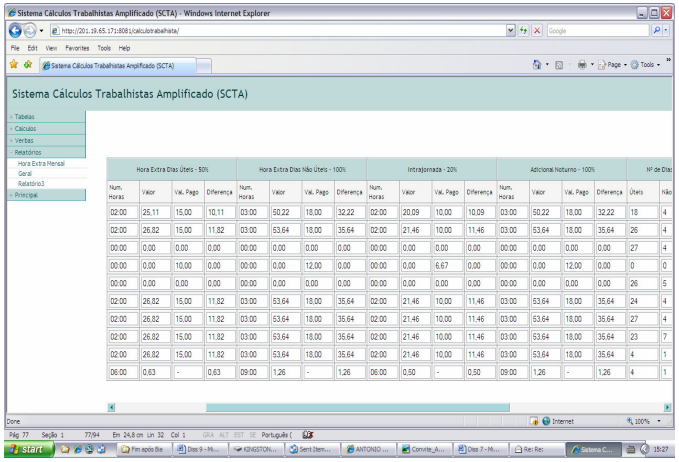


Figura 5.9 – Tela das Horas Extras.

- Demonstrativo dos Valores Apurados – um dos principais relatórios do sistema apresenta a remuneração do Reclamante, os valores dos pedidos deferidos e calculados mensalmente, suas respectivas bases de cálculo, reflexos e incidências.

- Demonstrativo da Contribuição Previdenciária – mostra um detalhamento da apuração do INSS deduzido do autor e o que é devido pela Reclamada mês a mês, com suas respectivas alíquotas.

- Juros e Correção Monetária – geram o valor total histórico devido de pedidos (V1) e de FGTS (V2) ao autor em cada mês, corrigidos por um fator mensal oriundo da tabela de "Valor da Correção Monetária", multiplicado também pelos juros simples do período (1% ao mês).

- IRRF e Conversão dos valores para IDTR – neste último relatório, são apresentados: o valor V1 subtraído do imposto de renda e convertido para

IDTR's (V3); valor do FGTS (V2) convertido também para IDTR's (V4). O valor de V3 mais V4 é o total devido ao cliente; o valor do imposto de renda que a Reclamada deverá recolher e o valor do INSS obtido com base no demonstrativo "Demonstrativo da Contribuição Previdenciária".

6. Resultados Experimentais

O objetivo deste capítulo é apresentar os estudos de casos realizados. Também serão apresentados os resultados encontrados através dos vários experimentos durante as etapas do sistema demonstrando que os resultados encontrados são muito bons para o objeto da aplicação

6.1. Coleção de documentos para escolha do melhor classificador

O primeiro passo é a geração dos arquivos de treinamento e teste. Para isto foram utilizadas 104 "bolsas de palavras" (BP), de quatro classes (pedidos) diferentes. Cada BP possui aproximadamente entre mil e duas mil palavras. A tabela 6.1 apresenta um quadro resumo com a quantidade de BP por classe assim como a seleção para treinamento e teste. Ressalta-se que foram geradas cinco análises com grupos de arquivos diferentes para a validação cruzada que foi empregada utilizando-se cinco subamostragens aleatórias (random subsampling).

Tabela 6.1 – distribuição dos arquivos gerados

Classe	treinamento	teste	total	% tes/total
Alimentação	12	5	17	**29%**
equiparação salarial	11	5	16	**31%**
hora extra	25	12	37	**32%**
honorário advocatício	23	11	34	**32%**
Total	**71**	**33**	**104**	

6.2. Processamento para escolha do melhor classificador a ser utilizado na Mineração de Texto (MT)

Nesta seção serão apresentados os diversos classificadores utilizados e seus respectivos resultados com objetivo de escolher o melhor a ser usado para classificação dos novos documentos. Para a escolha dos classificadores utilizaram-se os índices de precisão, cobertura e medida F definidas no capítulo 2. Quanto maior forem seus valores, melhores serão os resultados. Na escolha do melhor por classe, foi selecionado

o que tivesse o maior valor da medida F que procura balancear a relação entre os índices de precisão e cobertura.

6.2.1. Método *Naive Bayes* (NB)

Para análise dos resultados foram utilizados os parâmetros do classificador abaixo descritos:

a) Quantidade de palavras do dicionário de dados - representa a quantidade de palavras que serão selecionadas e utilizadas dentro dos documentos para as análises estatísticas. Quanto maior este parâmetro, mais palavras serão utilizadas para análise no aplicativo;
b) Mínimo de freqüência – significa a quantidade mínima de ocorrências de uma palavra para ser incluída no dicionário de dados. Quanto maior for este parâmetro mais vezes uma palavra terá que ocorrer no texto para ser selecionada para o dicionário de dados;
c) Limiar de probabilidade – reflete o valor de probabilidade para se classificar um documento. Possui valor *default* igual a 0,5.
d) Limiar de rejeição – valor limite que deve ser excedido para classificar um documento. O valor *default* é 0,5. Quanto maior, mais difícil de classificar um documento.

Depois de alguns testes de variabilidade, foi realizada uma analise de sensibilidade para a escolha do melhor conjunto de parâmetros empregando-se os valores apresentados na tabela 6.2. Ressalta-se que os parâmetros Limiar de probabilidade e rejeição, foram testados com valor muito baixo (0.000001) e muito alto (0.99999), pois com outros valores intermediários não apresentavam variação nos resultados diferentes dos parâmetros com 0.1, 0.5 e 0.9.

Tabela 6.2 – Valores utilizados para análise de sensibilidade com o classificador NB

Parâmetros	Valores
Stemming	Com
Stopwords	Com
Quantidade de palavras do dicionário	30, 50, 500
Freqüência mínima de palavras	1, 50 100
Limiar de probabilidade	0.1, 0.9, 0.000001, 0.99999
Limiar de rejeição	0.1, 0.5, 0.000001, 0.99999

O anexo B apresenta um arquivo *batch,* com os principais comandos, para executar este classificador. A análise de sensibilidade destes parâmetros foi feita através de 145 alternativas de combinações dos parâmetros, descritas na tabela 6.2, gerando 725 resultados diferentes devido às cinco subamostragens. Os resultados completos com as médias das subamostragens encontram-se no anexo C. Os melhores resultados obtidos por classe são demonstrados na tabela 6.3 e os respectivos valores dos parâmetros estão apresentados na tabela 6.4.

Tabela 6.3 – Melhores resultados encontrados por classe para o classificador NB

Classe	Precisão (%)	Cobertura (%)	Medida F p/classe (%)	Medida F classificador (%)	Número registro teste
Alimentação	87,95	83,98	85,92	85,53	103
Equiparação	88,94	87,88	88,40	88,38	105
Horas Extras	87,77	85,92	86,84	78,76	115
Honorários	94,84	95,78	95,31	90,67	57
média			89,12	85,83	98

Tabela 6.4 – Parâmetros utilizados nos melhores resultados encontrados por classe para o classificador NB

Classe	Quantidade palavras	Freqüência Minima	Limiar probabilidade	Limiar rejeição
Alimentação	500	1	0,90	0,000001
Equiparação	500	1	0,000001	0,10
Horas Extras	500	50	0,10	0,000001
Honorários	50	1	0,000001	0,10
Medida F classificador	500	1	0,10	0,50

A conclusão obtida a partir da análise de sensibilidade é que a classificação para cada classe ocorre com um conjunto de diferentes valores dos parâmetros o que demonstra ser possível calibrar cada parâmetro por classe com objetivo de obter o melhor desempenho.

6.2.2. Método Linear por Ordenação

Os resultados foram analisados variando-se os parâmetros descritos abaixo, após análise prévia da variabilidade dos mesmos:

a) Quantidade de palavras do dicionário de dados e Mínimo de freqüência – são os mesmos parâmetros definidos no método Naive Bayes;
b) Limiar de decisão – Controla o *tradeoff* entre precisão e cobertura. Possui valor *default* igual a 0,3.
c) Lambda - controla o tamanho do espaço de procura. Valor *default* é de 0.001.
d) Taxa de aprendizado – tem valor *default* igual a 0.25.
e) Tipo de característica – define a forma de armazenamento dos termos. Default é *tf (term frequency)*.

Depois de alguns testes de variabilidade, foram utilizados os valores dos parâmetros apresentados na tabela 6.5. Ressalta-se que o parâmetro Lambda foi testado com valor muito baixo (0.00001) devido ao fato de que com outros valores não apresentava variação significativa.

Tabela 6.5 – Valores utilizados para análise de sensibilidade com o classificador Linear

Parâmetros	Valores
Stemming	Com
Stopwords	Com
Quantidade de palavras do dicionário	30, 50, 500
Freqüência mínima de palavras	1, 50, 100
Limiar de decisão	-0.8, 0.2, 0.8
Lambda	0.01, 0.9, 0.00001
Taxa de aprendizado	0.25, 0.9, 0.01
Tipo da característica	Binary, tf, tf*idf

O anexo D apresenta um dos arquivos *batch,* com os principais comandos, para executar este classificador. A análise de sensibilidade destes parâmetros foi feita através de 323 alternativas de combinações dos parâmetros, descritas na tabela 6.5 obtidas para cada subamostragem. Os resultados completos com as médias das subamostragens encontram-se no anexo E. Os principais resultados para cada classe estão demonstrados na tabela 6.6 e os respectivos valores dos parâmetros estão apresentados na tabela 6.7.

Tabela 6.6 – Melhores resultados encontrados por classe para o classificador Linear

Classe	Precisão (%)	Cobertura (%)	Medida F p/classe (%)	Medida F classificador (%)	Número registro teste
Alimentação	78,95	96,57	86,88	86,88	110
Equiparação	96,90	94,77	95,82	95,44	109
Horas Extras	92,21	96,49	94,30	90,90	275
Honorários	88,98	84,97	86,92	86,66	111
média			90,98	89,97	110

Tabela 6.7 – Parâmetros utilizados nos melhores resultados encontrados por classe para o classificador Linear

Classe	Quantidade palavras	Frequência Mínima	Limiar de decisão	Lambda	Taxa aprend.	Tipo Caracter.
Alimentação	500	1	0,20	0,01	0,90	tf
Equiparação	500	1	0,20	0,01	0,25	tf*idf
Horas Extras	500	50	(0,80)	0,01	0,01	tf*idf
Honorários	500	1	0,20	0,01	0,90	tf*idf
Medida F classificador	500	1	0,20	0,01	0,90	tf

6.2.3. Método por Indução de Regras

Para análise dos resultados foram utilizados os parâmetros do classificador abaixo descritos:

a) Quantidade de palavras do dicionário de dados - representa a quantidade de palavras que serão selecionadas e utilizadas dentro dos documentos para as análises estatísticas. Quanto maior este parâmetro, mais palavras serão utilizadas para análise no aplicativo;
b) Limiar de Freqüência – significa a freqüência limite. Valor default = 1 (binário);
c) Tipo de Teste – define o tipo de teste nas regras. Possui valor *default* igual a 1.
d) Tradeoff entre precisão e cobertura – permite definir o limite entre precisão e cobertura. Default igual a 0.
e) Seleção – especifica como escolher o melhor conjunto de regras.

Depois de alguns testes de variabilidade, foram utilizados os valores dos parâmetros apresentados na tabela 6.8.

Tabela 6.8 – Valores utilizados para análise de sensibilidade c/o classificador p/Indução de Regras

Parâmetros	Valores
Stemming	Com
Stopwords	Com
Quantidade de palavras do dicionário	30, 50, 100
Limiar de freqüência	0, 1 e 2
Tipo de Teste	1 e 2
Tradeoff precisão x cobertura	0, 4 e 9
Seleção	0, 1 e 6

O anexo F apresenta um arquivo *batch,* com os principais comandos, para executar este classificador. A análise de sensibilidade destes parâmetros foi feita através de 104 alternativas de combinações dos parâmetros, descritas na tabela 6.8, gerando 520 resultados diferentes devido às cinco subamostragens. Os resultados completos com as médias das subamostragens encontram-se no anexo G. Os melhores resultados por classe são demonstrados na tabela 6.9 e os respectivos valores dos parâmetros estão apresentados na tabela 6.10.

Tabela 6.9 – Melhores resultados encontrados por classe para o classificador por Indução de Regras

Classe	Precisão (%)	Cobertura (%)	Medida F p/classe (%)	Medida F classificador (%)	Número registro teste
Alimentação	85,74	61,57	71,67	82,80	56
Equiparação	82,22	85,08	83,62	85,10	53
Horas Extras	85,05	85,47	85,26	75,80	31
Honorários	75,41	82,19	78,66	80,16	95
média			79,80	80,96	57

Tabela 6.10 – Parâmetros utilizados nos melhores resultados encontrados por classe para o classificador por Indução de Regras

Classe	Quantidade palavras	Limiar Frequencia	Tipo Teste	Tradeoff	Seleção
Alimentação	100	1	1	9	0
Equiparação	100	0	2	9	0
Horas Extras	50	1	2	0	0
Honorários	30	1	2	4	1
Medida F classificador	100	1	2	0	0

Tal como com os outros classificadores, a análise de sensibilidade demonstra que a classificação para cada classe ocorre com um conjunto de diferentes valores dos parâmetros o que demonstra ser possível calibrar cada parâmetro por classe com objetivo de obter o melhor desempenho.

6.2.4. Resumo dos resultados

A partir da análise dos três classificadores, concluímos que os melhores resultados para cada classe podem ser obtidos através de classificadores diferentes. A tabela 6.11 apresenta os melhores resultados por classe.

Tabela 6.11 – Melhores resultados encontrados por classe

Classe	Precisão (%)	Cobertura (%)	Medida F p/classe (%)	Classificador
Alimentação	78,95	96,57	86,88	linear
Equiparação	96,90	94,77	95,82	linear
Horas Extras	92,21	96,49	94,30	linear
Honorários	94,84	95,78	95,31	naive bayes

6.3. Processamento do Sistema Especialista (SE)

Após serem usadas as técnicas de MT para identificar as verbas deferidas e os respectivos parâmetros complementares, utiliza-se o SE, para cálculo dos valores que a empresa deve ao funcionário. A seguir, será apresentado o funcionamento do sistema, através de algumas das principais telas, iniciando-se pelo cadastramento nas tabelas, passando pela digitação da base de cálculo e quantidade de horas extras deferidas, finalizando com o valor devido pela empresa.

a) Cadastramento dos valores bases das tabelas – inicialmente, para o sistema operar, é necessário cadastramento de várias informações básicas em repositórios de dados tais como: Alíquotas do INSS, Histórico do Salário Mínimo, Tabela de Feriados nacionais e locais, entre outros. A figura 6.1 apresenta a tela de cadastramento das alíquotas.

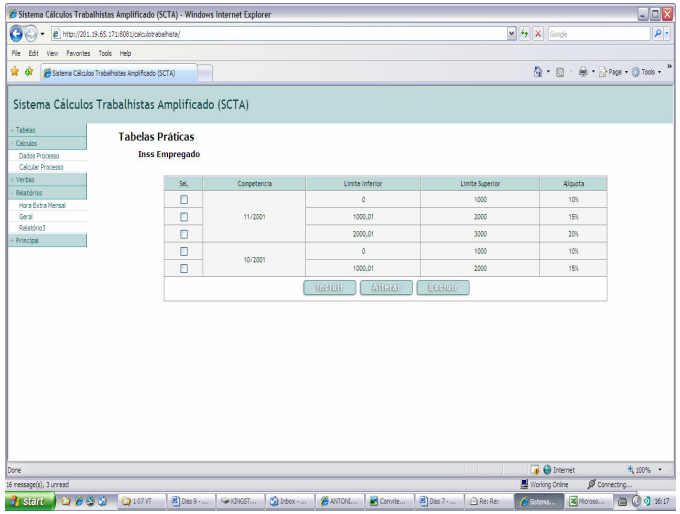


Figura 6.1 – Alíquotas de INSS

b) Base de cálculo – o próximo passo é a digitação dos dados do processo (nome do funcionário, nome da empresa, data de admissão, demissão, etc..) e também dos valores que compõem a base de cálculo da remuneração do Autor. Esses valores são normalmente compostos de salário, comissão, adicional por tempo de serviços, etc.. e podem ser obtidos externamente ao sistema. A figura 6.2 contém a tela de entrada de dados da base de cálculo do sistema.

c) Verbas – a seguir o sistema será alimentado com as verbas deferidas no processo e seus parâmetros adicionais. No exemplo em questão será assumido que a verba deferida foi hora extra, tendo o juiz definido a quantidade de 10 (dez) horas por mês em dias úteis e 20 (vinte) em dias não úteis. A figura 6.3 apresenta a tela de entrada de dados das horas extras.

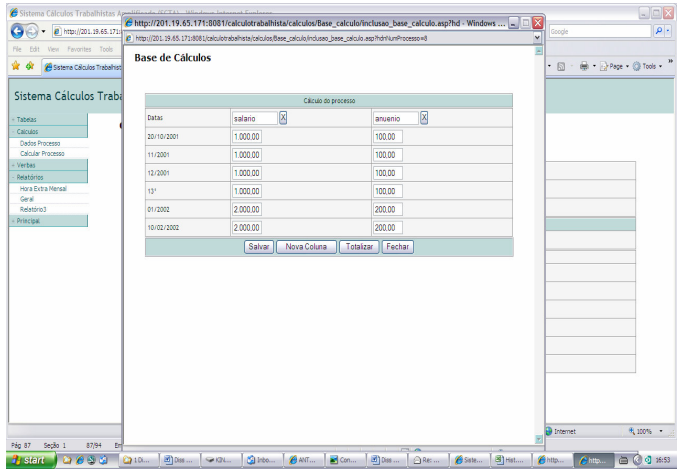


Figura 6.2 – Base de cálculo

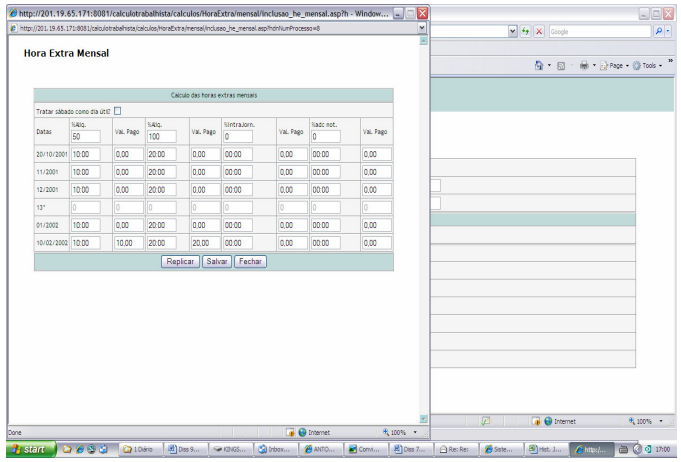


Figura 6.3 – Tela de Horas Extras

d) Resultado – ao final, nesta simulação, para estes salários e horas extras definidos, o SE apresenta na coluna “Total Geral”, última linha da figura 6.4, o valor histórico que a empresa deve ao empregado.

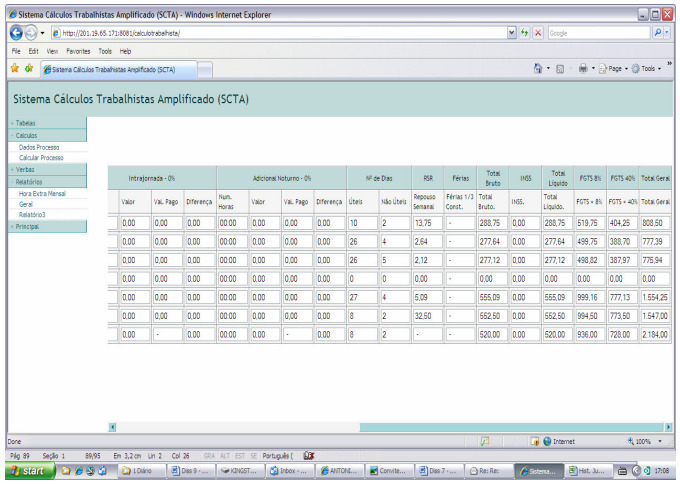


Figura 6.4 – Resultado Final

Ressalta-se que o sistema é bastante amigável, possibilitando não somente que peritos possam operá-los, como um advogado ou até mesmo um assistente. Busca também ser parametrizado ao máximo, para evitar manutenções que poderiam gerar instabilidade nos resultados.

7. Conclusão

A área jurídico-trabalhista, além de tratar grande volume de informação não estruturada (textos), também requer sempre, ao final das decisões dos Juízes, se favorável ao Autor da ação, o cálculo dos valores devidos pela empresa a estes, como também os impostos (INSS, IR). Logo, existe necessidade de se fazer um grande volume de cálculos. Atualmente, existem vários sistemas que implementam estes cálculos, mas dentro do que foi pesquisado neste trabalho, nenhum deles automatiza a interpretação das peças jurídicas (sentenças, embargos e acórdãos), todos se baseiam na leitura destas peças por um especialista e posteriormente digitação dos dados em um sistema convencional.

Por outro lado, o mecanismo desenvolvido neste trabalho visa, justamente, informatizar a parte relativa ao tratamento de texto, peças jurídicas, com um sistema especialista que irá calcular o valor final que a empresa deverá pagar ao cliente, ou seja, evitando a leitura destas peças e digitação, por um especialista.

Os resultados encontrados na mineração de texto foram satisfatórios, visto que foram encontrados valores para o principal indicador do desempenho (medida F) acima de 94% em todas as classes, com exceção da classe “alimentação” (84%). Ressalta-se que para atingir esta performance, foi testado cada um dos tipos de classificadores com várias calibragens de parâmetros diferentes. Ao final, conclui-se que cada classe pode ter melhor resultado com um método de classificação diferente.

O classificador de método Linear foi o que obteve o mais alto desempenho, tendo obtido o melhor resultado em três classes (equiparação, hora extra e honorário). O tipo *Naive Bayes* foi o melhor na classe alimentação. O método por Indução de Regras, em que pese, ser o que apresenta maior clareza nos resultados encontrados, por implementar lógicas conhecidas do ser humano, chegou aos piores resultados, não tendo sido o mais alto em nenhuma das classes, tendo ainda chegado a valores abaixo de 76%.

Analisando os resultados como um todo, concluímos que as "bolsas de palavras" (BP), apesar de terem uma grande quantidade de palavras, entre mil e duas mil, acarretando grandes dimensionalidades nos arquivos de características, por outro lado, favorecem a identificação das classes.

Para avaliar o impacto da grande quantidade de palavras das BP's desta aplicação, foram realizados alguns testes utilizando uma aplicação Java que implementa o algoritmo "K-vizinho mais próximo", com as medidas de distâncias: euclidiana, *manhattan, camberra e minimax*. Este tipo de algoritmo é tido como dos mais tradicionais e usados, mas requer grande capacidade de recursos de máquina. Durante os testes, o sistema não completava o processamento necessário, comprovando a dificuldade de se tratar o grande volume de informações manipuladas por esta aplicação.

7.1 Trabalhos Futuros

O módulo central do sistema, entre a mineração de textos e o sistema especialista, deve ser desenvolvido e implementado. Este módulo contempla as seguintes etapas:

- Identificação se o pedido foi deferido ou indeferido. Isto pode ser feito através de um programa de linguagem normal acoplado a um dicionário de dados (*thesaurus*), visto existirem várias palavras similares que tem o mesmo significado tal como: defere é igual a deferimento, dou seguimento, é devido, etc... e indefere é igual a indeferimento, nego seguimento, não é devido.

- Selecionar as incidências/reflexos que os pedidos estão gerando. Como exemplos podem ser citados: FGTS, 13 salários, férias, etc...

- Capturar alguns outros parâmetros do pedido, tal como o horário deferido em caso de horas extras, o percentual que define o adicional de insalubridade, entre outros.

- Extrair algumas outras informações relativas ao processo tal como: data da prescrição, nome do reclamante, nome da reclamada e outros.

Em que pese parecerem ser muitos itens, os mesmos correspondem a apenas 15% da aplicação no que tange a redução de erros e rapidez, ou seja, a implantação do que foi feito durante este trabalho equivale a aproximadamente 85%. Estes itens não são complexos de serem implantados, visto poderem ser desenvolvidos através de linguagem de programação convencional.

Ao final, depois de capturadas todas as informações necessárias, será gerado um arquivo de interface para o sistema especialista (SE) no seguinte *layout*:

Campo 1	**Campo 2**	**Campo 3**	**Campo 4**
tipo do pedido	d/i (deferido ou indeferido)	reflexo 1..... reflexo n	parâmetros 1 ... parâmetros n

Referências Bibliográficas

BASTOS, V. M., 2006, *Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

BELKIN, N. J., CROFT, W. B., 1992, "Information filtering and information retrieval: Two sides of the same coin?", *Communications of the ACM*, v. 35, n. 12, pp. 29-38.

BOLBOACÃ, S., JŽNTSCHI, L., 2006, "Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds", *Leonardo Journal of Sciences*, v. 5, n. 9, pp. 179-200.

BRAY, T, PAOLI, J., SPERBERG-McQUEEN, C. M., MALER, E., 2000, "Extensible Markup Language (XML) 1.0 (Second Edition) – W3C Recommendation 6", disponível no site www.w3.org/TR/2000/REC-xml-20001006, último acesso em 26 de março de 2007.

COHEN, W., HIRSH, H., 1998, "Joins that Generalize: Text Classification Using WHIRL". In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 169-173, New York, Aug.

COHEN, W., SINGER, Y., 1999, "Context-Sensitive Learning Methods for Text Categorization", *ACM Transactions on Information Systems*, v. 17, n. 2 (Ap.), pp. 141-173.

COWIE, J., LEHNERT, W., 1996, "Information extraction", *Communications of the ACM*, v. 39, n. 1, pp. 80-91.

DAGAN, I., KAROV, Y., ROTH, D., 1997, "Mistake-driven learning in text categorization". In: *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*, pp. 55-63, Providence, Jun.

DOYLE, J., 1996, "Strategic Directions in Artificial Intelligence", *ACM Computer surveys*, v. 28, n. 4, pp. 653-669.

FELFERING, A., KOSTYANTYN, S., 2006, "Debugging user interface descriptions of knowledge-based recommender applications". In: *Proceding of the 11th International Conference on Intelligent user interfaces*, pp. 234-241, Sydney, Jan.-Feb.

FUHR, N., HARTMANN, S., LUSTIG, G., 1991, "AIR/X – a Rule-Based Multistage Indexing System for Large Subject Fields". In: *Proceedings of RIAO-1991 3rd International Conference: Recherche d'Information Assistee par Ordinateur*, pp. 606-623, Barcelona, April.

FULLAM, K., PARK, J., 2002, "Improvements for Scalable and Accurate Plagiarism Detection in Digital Documents", University of Texas at Austin, site https://webspace.utexas.edu/fullamkk/pdf/DataMiningReport.pdf, último acesso em 02 de maio de 2007.

GROOTHUIS, M. M., SVENSSON, J. S., 2000, "Expert system support and juridical quality". In: *Proceedings of Legal Knowledge and Information Systems*, pp. 1-10, Amsterdam.

HAN, J., KAMBER, M., 2001, *Data Mining: Concepts and Techniques*. 1ª ed., San Francisco, Morgan Kaufmann Publishers.

HART, A., 1986, *A Knowledge Acquisition for Expert Systems*. 2ª ed., New York, Mc Graw-Hill.

HINGORANEY, R., 1994, "Putting expert systems to work", *Chemical Engineering*, v. 101, n. 1 (Jan.), pp. 121-124.

JAIN, A. K., MURTY, M. N., FLYNN, P. J., 1999, "Data Clustering: a Review", *ACM Computing Surveys*, v. 31, n. 3, pp. 264-323.

JIZBA, R., "Measuring Search Effectiveness", Creighton University Health Sciences Library and Learning Resources Center, Nebraska. Disponível em http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html, último acesso em 12 de abril de 2007.

JOACHIMS, T., 1998, *Making large scale SVM learning practical*, LS8 Report 24, University of Dortmund Fachbereich Informatik Lehrstuhl.

JOACHIMS, T., 1998, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features". In: *Proceedings of the 10th European Conference on Machine Learning*, pp. 137-142, Chemnitz, April.

JOACHIMS, T., 2002, *Learning to Classify Text Using Support Vector Machines, Methods, Theory and Algorithms*. 1ª ed., Norwell, Kluwer Academic Publishers.

KHAN, M., DING, Q., PERRIZO, W., 2002, "K-nearest neighbor classification spatial data streams Using P-trees". In: *Proceedings of the PAKDD, Pacific-Asia Conference on Knowledge Discoveryand Data Mining*, pp. 517-118, Taipei, May.

KONGTHON, A., 2004, *A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management*. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, USA.

KRAAIJ, W., POHLMANN, R., 1996, "Viewing stemming as recall enhancement". In: *Annual ACM Conference on Research and Development in Information Retrieval – Proceedings of the 19th annual international SIGIR*, pp. 40-48, Zurich, Aug.

KRUSE R., BORGELT C., 2003, "Information Mining", *International Journal of Approximate Reasoning*, v. 32, n. 2, pp. 63-65.

LANDAUER, T. K., DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., 1990, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, v. 41, n. 6, pp. 391-407.

LANDAUER, C., 1990, "Correctness principles for rule-based expert systems", *Expert Systems with Applications*, v. 1, n. 3, pp. 291-316.

LI, Y. H., JAIN, A. K., 1998, "Classification of text documents", *Computer Journal*, v. 41, n. 8, pp. 537-546.

LOPES, M. C., 2004, *Mineração de Dados Textuais utilizando técnicas de Clustering para o Idioma Português*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MAK, B., BLANNING, R., 2003, "A logic-based approach to rule induction in expert systems". *Expert Systems*, v. 20, n. 3 (Jul.), pp. 123-162.

McCALLUM, A., NIGAM, K., 1998, "A Comparison of Event Models for Naive Bayes Text Classification". In: *AAAI-98 – Workshop on Learning for Text Categorization*, pp. 41-48, Madison, July.

MITCHELL, T. M., 1997, *Machine Learning*. 1ª ed., New York, McGraw-Hill.

MOYNIHAN, G. P., SUKI, A., FONSECA, D. J, 2006, "An expert system for the selection of software design patterns", *Expert Systems*, v. 23, n. 1 (Feb.), pp. 39-52.

O'CALLANGHAN, T. A., POPPLE, J., McCREATHET, E., 2003, *Building and Testing the SHYSTER-MYCIN Hybrid Legal Expert System*, Technical Report TR-CS-03-01, Australian National University, Canberra.

ORENGO, V. M., HUYCK, C., 2001, "A Stemming Algorithm for the Portuguese Language". In: *8th International Symposium on String Processing and Information Retrieval*, pp. 183-193, Laguna de San Raphael, Nov.

RADEV, D., FAN, W., ZANG, Z., 2001, "Webinessence: A Personalized Web-Based Multi-Document Summarization and Recommendation System". In: *NAACL Workshop on Automatic Summarization*, Pittsburgh.

RAHAL, I., PERRIZO, W., 2004, "An optimized Approach for KNN Text Categorization using P-tree". In: *Proceedings of the 2004 ACM Symposium on Applied computing*, pp. 613-617, Nicosia, Mar.

RISH, I., 2001, "An empirical study of the naive Bayes classifier". In: *Proceedings of IJCAI-01 Workshop on Empirical Methods*. Disponível em http://www.cc.gatech.edu/fac/Charles.Isbell/classes/reading/papers/Rish.pdf, último acesso em 02 de maio de 2007.

ROCCHIO, J., 1971, "Relevance Feedback in information retrieval". In: Salton, G. (ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing*, Chapter 14, New Jersey, USA, Prentice-Hall Inc.

RUSSEL, S., NORVIG, P., 2004, *Inteligência Artificial*. 2ª ed., Rio de Janeiro, Elsevier.

SAGHEB, M., 2006, "The design process of expert systems development: some concerns", *Expert Systems*, v. 23, n. 2 (May), pp. 116-125.

SALTON, G., McGILL, M. J., 1983, *Introduction to modern information retrieval*. New York, McGraw- Hill.

SALTON, G., 1989, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Massachusetts, Addison-Wesley.

STATISTICA TEXT MINER, 2005, site disponível em http://www.statsoft.com/products/textminer.html, último acesso em 26 de março de 2007.

STEINBACH, M., KARYPIS, G., KUMAR, V., 2000, "Comparison of Document Clustering Techniques". In: *KDD Workshop on Text Mining*, Boston, August. Disponível em http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf, último acesso em 02 de maio de 2007.

TAN, A., 1999, "Text Mining: The state of the art and the challenges". In: *Proceedings of the Pacific-Asian Conference on Knowledge Discovery and Data Mining*, Beijing, April.

TICOM, A., 2007, "Text Mining and Expert System applied in Labor Laws", In: 7th International Conference on Intelligent Systems Design and Applications, Rio de Janeiro, Brasil.

VAPNIK, V., 1999, *The Nature of Statistical Learning Theory*. 2ª ed., New York, Springer-Verlag.

WATERMAN, D., 1986, *A Guide to Expert System*, Addison-Wesley Publishing Company.

WEISS, S. W., INDURKHYA, N., ZHANG, T., DAMERAU, F. J., 2004, *Text Mining: Predictive Methods for Analizing Unstructured Information*. New York, Springer.

XAVIER, A. E., "The Hyperbolic Smoothing Clustering Method", Dept. of Systems Engineering and Computer Science. Disponível em http://cronos.cos.ufrj.br/publicacoes/reltec/es67405.pdf, último acesso em 04 de maio de 2007.

ZHANG, J., YANG, Y, 2003, "Robustness of regularized linear classification methods in text categorization". In: *Annual ACM Conference on Research and Development in Information Retrieval*, pp. 190-197, Toronto.

ZHANG, T., OLES, F. J., 2001, "Text categorization based on regularized linear classification methods", *Information Retrieval*, v. 4, n. 1 (Ap.), pp. 5-31.

ANEXO

A) Lista de Stop words

De	ele	num	delas	houvéramos	seremos	qualquer
assim	das	nem	esta	haja	serão	cada
afim	tem	suas	estes	hajamos	seria	após
agora	à	meu	estas	hajam	seríamos	durante
onde	seu	às	aquele	houvesse	seriam	entanto
outro	sua	minha	aquela	houvéssemos	tenho	sempre
outros	ou	têm	aqueles	houvessem	tem	menos
ainda	ser	numa	aquelas	houver	temos	mais
A	quando	pelos	isto	houvermos	tém	caso
o	muito	elas	aquilo	houverem	tinha	segundo
que	há	havia	estou	houverei	tínhamos	àqueles
vario	nos	seja	está	houverá	tinham	destas
varios	já	qual	estamos	houveremos	tive	todos
vário	está	será	estão	houverão	teve	
vários	eu	nós	estive	houveria	tivemos	
e	também	tenho	esteve	houveríamos	tiveram	
do	só	lhe	estivemos	houveriam	tivera	
da	pelo	deles	estiveram	sou	tivéramos	
uns	pela	essas	estava	somos	tenha	
em	até	esses	estávamos	são	tenhamos	
um	isso	pelas	estavam	era	tenham	
para	ela	este	estivera	éramos	tivesse	
é	entre	fosse	estivéramos	eram	tivéssemos	
antes	era	dele	esteja	fui	tivessem	
anti	depois	tu	estejamos	foi	tiver	
com	sem	te	estejam	fomos	tivermos	
não	mesmo	vocês	estivesse	foram	tiverem	
uma	aos	vos	estivéssemos	fora	terei	
os	ter	lhes	estivessem	fôramos	terá	
no	seus	meus	estiver	seja	teremos	
se	quem	minhas	estivermos	sejamos	terão	
Na	nas	teu	estiverem	sejam	teria	
por	me	tua	hei	fosse	teríamos	
mais	esse	teus	há	fôssemos	teriam	
As	eles	tuas	havemos	fossem	porém	
dos	estão	nosso	hão	for	todavia	
como	você	nossa	houve	formos	entretanto	
mas	tinha	nossos	houvemos	forem	contudo	
foi	foram	nossas	houveram	serei	quer	
Ao	essa	dela	houvera	será	quais	

B) Programa Batch com os principais comando para executar o classificador Naive Bayes

PRIMEIRA EXECUÇÃO – C/ARQUIVO DE TREINAMENTO

```
java mkdict 1000 global.dit

java vectorize alimen ve.vec
java nbayes nb1ctr.wts

java vectorize equipa ve.vec
java nbayes nb2ctr.wts

java vectorize hext ve.vec
java nbayes nb3ctr.wts

java vectorize honora ve.vec
java nbayes nb4ctr.wts
```

SEGUNDA EXECUÇÃO – C/ARQUIVO DE TESTE

```
java vectorize alimen ve.vec
java testnbayes nb1ctr.wts nb1ctepo.txt nb1ctene.txt

java vectorize equipa ve.vec
java testnbayes nb2ctr.wts nb2ctepo.txt nb2ctene.txt

java vectorize hext ve.vec
java testnbayes nb3ctr.wts nb3ctepo.txt nb3ctene.txt

java vectorize honora ve.vec
java testnbayes nb4ctr.wts nb4ctepo.txt nb4ctene.txt
```

C) Resultados encontrados com o classificador Naive Bayes

					Alimentação			Equiparação			Horas Extras			Honorários			
Num reg	Num Pal	Freq mín	Lim Prob	Lim Rej	prec	cob	med F	prec	cob	med F	prec	cob	med F	prec	cob	med F	med F media
1	30	1	0,1	0,1	27,27	23,86	25,45	88,27	87,49	87,88	87,05	85,48	86,26	94,98	95,02	95,00	73,65
2	30	1	0,1	0,5	27,27	23,29	25,12	88,77	87,13	87,94	87,97	85,45	86,69	94,39	95,65	95,02	73,69
3	30	1	0,1	0,000001	27,95	23,79	25,70	88,23	87,53	87,88	87,16	85,27	86,20	94,60	95,57	95,08	73,72
4	30	1	0,1	0,99999	87,75	5,59	10,51	88,40	27,32	41,75	87,41	8,95	16,23	94,89	68,47	79,54	37,01
5	30	1	0,9	0,1	87,42	5,86	10,99	88,10	87,39	87,75	87,23	68,76	76,90	94,38	86,79	90,43	66,52
6	30	1	0,9	0,5	87,26	6,90	12,79	88,12	87,20	87,66	87,73	68,93	77,20	94,18	86,18	90,00	66,91
7	30	1	0,9	0,000001	87,31	5,72	10,75	88,35	87,66	88,00	87,95	68,94	77,29	94,07	86,35	90,04	66,52
8	30	1	0,9	0,99999	87,11	3,50	6,73	88,94	27,26	41,73	87,09	8,57	15,60	94,54	68,55	79,47	35,88
9	30	1	0,000001	0,1	18,05	83,82	29,70	88,19	87,31	87,75	79,48	85,13	82,21	94,36	95,56	94,95	73,65
10	30	1	0,000001	0,5	18,52	83,64	30,32	88,03	87,84	87,93	79,38	85,61	82,37	94,45	95,09	94,77	73,85
11	30	1	0,000001	0,000001	18,07	83,61	29,72	88,60	87,54	88,07	79,20	85,89	82,41	94,84	95,40	95,12	73,83
12	30	1	0,000001	0,99999	5,84	5,53	5,68	88,26	27,35	41,75	87,30	8,85	16,07	94,11	68,30	79,15	35,66
13	30	1	0,99999	0,1	87,87	6,26	11,68	88,18	27,93	42,43	87,03	8,81	16,00	94,93	68,31	79,45	37,39
14	30	1	0,99999	0,5	87,45	0,62	1,24	88,72	27,23	41,67	87,72	8,91	16,18	94,46	68,11	79,15	34,56
15	30	1	0,99999	0,000001	87,89	1,25	2,47	88,80	27,91	42,47	87,39	8,87	16,11	94,73	68,49	79,50	35,14
16	30	1	0,99999	0,99999	87,59	3,54	6,80	88,63	42,34	57,30	87,96	37,74	52,82	94,17	73,71	82,69	49,90
17	30	50	0,1	0,1	23,60	63,62	34,43	17,27	67,10	27,47	87,13	85,23	86,17	57,21	86,45	68,85	54,23
18	30	50	0,1	0,5	23,88	63,08	34,65	17,64	67,77	27,99	87,03	85,27	86,14	57,62	86,50	69,16	54,49
19	30	50	0,1	0,000001	23,28	63,02	34,00	17,44	67,56	27,73	87,71	85,56	86,62	57,06	86,17	68,66	54,25
20	30	50	0,1	0,99999	87,79	4,44	8,45	88,89	3,31	6,39	87,45	0,20	0,40	94,28	2,97	5,76	5,25
21	30	50	0,9	0,1	87,15	5,53	10,40	88,72	6,40	11,95	87,78	68,22	76,77	61,75	68,73	65,05	41,04
22	30	50	0,9	0,5	87,95	5,48	10,31	88,50	3,29	6,35	87,88	68,64	77,08	61,86	68,03	64,80	39,63
23	30	50	0,9	0,000001	87,16	6,93	12,84	88,03	6,86	12,73	87,08	68,55	76,71	61,19	68,63	64,70	41,74
24	30	50	0,9	0,99999	87,94	3,69	7,07	88,18	0,62	1,24	87,05	4,28	8,15	94,62	5,49	10,38	6,71
25	30	50	0,000001	0,1	15,90	83,53	26,72	15,43	87,95	26,26	79,14	85,76	82,32	40,17	95,72	56,59	47,97
26	30	50	0,000001	0,5	15,79	83,86	26,57	15,95	87,09	26,97	79,86	85,08	82,39	40,88	95,46	57,24	48,29
27	30	50	0,000001	0,000001	15,40	83,48	26,00	15,80	87,72	26,78	79,25	85,67	82,34	40,58	95,45	56,95	48,02
28	30	50	0,000001	0,99999	87,53	6,46	12,03	88,31	1,94	3,80	87,33	0,13	0,26	2,37	0,18	0,33	4,10
29	30	50	0,99999	0,1	87,41	2,09	4,09	88,55	1,04	2,05	87,96	2,49	4,85	94,79	4,94	9,39	5,09
30	30	50	0,99999	0,5	87,80	1,13	2,23	88,04	6,61	12,29	87,37	5,10	9,63	94,02	1,91	3,75	6,97
31	30	50	0,99999	0,000001	87,18	4,72	8,95	88,10	2,12	4,13	87,90	0,02	0,03	94,58	0,93	1,84	3,74
32	30	50	0,99999	0,99999	87,69	5,90	11,20	88,99	1,75	3,44	87,05	0,30	0,60	94,40	2,65	5,16	5,10
33	30	100	0,1	0,1	12,23	83,98	21,35	20,09	87,99	31,02	79,63	85,29	82,36	49,20	95,23	64,88	49,90
34	30	100	0,1	0,5	12,79	83,05	22,16	20,66	67,42	31,63	79,32	85,51	82,30	49,33	95,39	65,03	50,28
35	30	100	0,1	0,000001	12,15	83,68	21,22	20,24	67,11	31,10	79,49	85,42	82,35	49,02	95,85	64,87	49,88
36	30	100	0,1	0,99999	87,38	0,68	1,36	88,60	6,21	11,61	87,22	1,53	3,00	94,99	4,68	8,92	6,22
37	30	100	0,9	0,1	87,52	6,40	11,93	88,50	5,79	10,87	87,18	6,55	12,18	94,78	4,37	8,35	10,83
38	30	100	0,9	0,5	87,65	5,40	10,17	88,02	6,32	11,79	88,00	5,27	9,94	94,03	0,02	0,04	7,98
39	30	100	0,9	0,000001	87,64	0,28	0,55	88,36	0,32	0,63	87,58	6,73	12,51	95,00	6,87	12,81	6,62
40	30	100	0,9	0,99999	87,01	1,17	2,31	88,29	4,44	8,45	87,66	0,53	1,05	94,66	5,20	9,85	5,42
41	30	100	0,000001	0,1	15,26	83,08	25,78	15,83	87,86	26,83	79,31	85,94	82,49	27,84	95,82	43,14	44,56
42	30	100	0,000001	0,5	15,18	83,94	25,71	15,08	87,19	25,71	79,96	85,85	82,80	27,49	95,49	42,68	44,23
43	30	100	0,000001	0,000001	15,36	83,79	25,96	15,00	87,85	25,63	79,55	85,80	82,56	27,33	95,29	42,48	44,16
44	30	100	0,000001	0,99999	87,93	1,06	2,09	88,85	3,95	7,56	87,26	3,35	6,45	94,69	5,35	10,13	6,56
45	30	100	0,99999	0,1	87,52	5,46	10,28	88,88	5,86	10,99	87,31	2,15	4,20	94,75	5,21	9,88	8,84
46	30	100	0,99999	0,5	87,09	5,53	10,40	88,45	2,07	4,04	87,67	4,38	8,35	94,98	5,08	9,64	8,11
47	30	100	0,99999	0,000001	87,36	5,46	10,27	88,68	3,16	6,11	87,48	6,83	12,68	94,35	5,18	9,82	9,72
48	30	100	0,99999	0,99999	87,18	5,64	10,60	88,78	6,23	11,64	87,82	5,25	9,90	94,34	0,23	0,46	8,15
49	50	1	0,1	0,1	87,28	43,84	58,36	88,64	87,61	88,12	79,46	77,29	78,36	94,78	86,80	90,62	78,87
50	50	1	0,1	0,5	87,66	43,53	58,18	88,94	87,61	88,27	79,48	77,57	78,51	94,54	86,28	90,22	78,79
51	50	1	0,1	0,000001	87,25	43,94	58,45	88,76	87,34	88,04	79,53	77,71	78,61	94,97	86,19	90,37	78,87
52	50	1	0,1	0,99999	87,25	3,44	6,62	88,65	27,02	41,42	87,97	68,55	77,06	94,11	77,06	84,74	52,46
53	50	1	0,9	0,1	87,75	23,43	36,98	88,86	87,81	88,33	79,27	77,23	78,24	94,58	77,88	85,42	72,24
54	50	1	0,9	0,5	87,02	23,33	36,79	88,09	87,72	87,90	79,94	77,38	78,64	94,79	77,60	85,34	72,17
55	50	1	0,9	0,000001	87,94	23,38	36,94	88,76	87,55	88,15	79,22	77,12	78,16	94,47	77,24	84,99	72,06
56	50	1	0,9	0,99999	87,23	1,28	2,52	88,32	27,35	41,76	87,89	68,99	77,30	94,57	77,85	85,40	51,75
57	50	1	0,000001	0,1	18,60	83,96	30,45	88,75	87,78	88,26	79,16	85,35	82,14	94,84	95,78	95,31	74,04
58	50	1	0,000001	0,5	18,18	83,29	29,85	88,29	87,97	88,13	79,04	85,00	81,91	94,50	95,62	95,06	73,74
59	50	1	0,000001	0,000001	18,40	83,30	30,15	88,77	87,67	88,22	79,94	85,78	82,76	94,23	95,53	94,88	74,00
60	50	1	0,000001	0,99999	2,71	1,95	2,27	88,61	27,24	41,67	87,52	68,47	76,83	94,07	86,49	90,12	52,72
61	50	1	0,99999	0,1	87,72	3,81	7,31	88,92	27,56	42,08	87,64	68,52	76,91	94,82	77,41	85,24	52,88
62	50	1	0,99999	0,5	87,48	5,18	9,78	88,50	27,99	42,53	87,68	68,69	77,03	94,25	77,48	85,04	53,60
63	50	1	0,99999	0,000001	87,86	1,92	3,76	88,94	27,14	41,59	87,72	68,38	76,85	94,37	77,16	84,90	51,78
64	50	1	0,99999	0,99999	87,74	4,49	8,54	88,71	27,09	41,51	87,18	68,57	76,76	94,90	77,47	85,31	53,03
65	50	50	0,1	0,1	23,82	63,48	34,65	17,02	67,07	27,16	87,69	85,93	86,80	57,16	86,48	68,83	54,36
66	50	50	0,1	0,5	23,11	63,77	33,93	17,29	67,07	27,50	87,08	85,25	86,16	57,68	86,26	69,14	54,18

67	50	50	0,1	0,000001	23,74	63,88	34,62	17,42	67,65	27,71	87,20	85,40	86,29	57,58	86,99	69,30	54,48
68	50	50	0,1	0,99999	87,52	0,09	0,18	88,69	5,55	10,44	87,13	4,11	7,84	94,19	2,42	4,71	5,79
69	50	50	0,9	0,1	87,37	3,16	6,10	88,34	5,51	10,37	87,31	68,11	76,52	61,33	68,48	64,70	39,42
70	50	50	0,9	0,5	87,42	5,63	10,58	88,49	3,19	6,16	87,07	68,31	76,56	61,48	68,68	64,88	39,54
71	50	50	0,9	0,000001	87,93	6,96	12,90	88,29	3,13	6,05	87,83	68,53	76,99	61,01	68,10	64,36	40,07
72	50	50	0,9	0,99999	87,05	6,50	12,09	88,62	5,66	10,64	87,54	0,49	0,98	94,33	6,65	12,42	9,03
73	50	50	0,000001	0,1	15,10	83,83	25,59	15,02	87,44	25,63	79,71	85,08	82,31	40,15	95,30	56,49	47,50
74	50	50	0,000001	0,5	15,44	83,88	26,08	15,13	87,29	25,79	79,18	85,12	82,04	40,16	95,03	56,46	47,59
75	50	50	0,000001	0,000001	15,21	83,06	25,71	15,71	87,68	26,65	79,54	85,89	82,59	40,98	95,33	57,32	48,07
76	50	50	0,000001	0,99999	87,88	5,29	9,99	88,91	3,03	5,87	87,33	2,08	4,06	2,45	3,15	2,76	5,67
77	50	50	0,99999	0,1	87,43	4,35	8,29	88,31	3,97	7,61	87,45	6,38	11,90	94,52	5,09	9,66	9,36
78	50	50	0,99999	0,5	87,65	5,85	10,96	88,12	4,33	8,25	87,07	2,02	3,95	94,48	1,11	2,20	6,34
79	50	50	0,99999	0,000001	87,32	2,71	5,26	88,05	3,92	7,51	87,30	0,80	1,58	94,54	1,80	3,54	4,47
80	50	50	0,99999	0,99999	87,90	5,40	10,17	88,32	3,62	6,96	87,15	6,75	12,53	94,18	6,88	12,83	10,62
81	50	100	0,1	0,1	12,28	83,17	21,40	20,54	67,54	31,50	79,02	85,48	82,12	49,66	95,69	65,39	50,10
82	50	100	0,1	0,5	12,80	83,77	22,20	20,93	67,82	31,99	79,38	85,30	82,23	49,45	95,91	65,25	50,42
83	50	100	0,1	0,000001	12,18	83,71	21,26	20,76	67,33	31,73	80,00	85,05	82,45	49,20	95,97	65,06	50,12
84	50	100	0,1	0,99999	87,79	5,62	10,56	88,80	6,93	12,86	87,73	4,59	8,72	94,66	5,93	11,16	10,83
85	50	100	0,9	0,1	87,96	1,05	2,08	88,93	6,33	11,82	87,09	5,74	10,77	94,18	6,32	11,86	9,13
86	50	100	0,9	0,5	87,06	0,67	1,33	88,38	2,43	4,74	87,48	0,70	1,40	94,99	2,72	5,28	3,19
87	50	100	0,9	0,000001	87,97	4,46	8,50	88,76	4,63	8,80	87,61	6,25	11,66	94,31	4,59	8,76	9,43
88	50	100	0,9	0,99999	87,87	6,14	11,48	88,52	1,02	2,02	87,14	3,23	6,23	94,00	3,25	6,28	6,50
89	50	100	0,000001	0,1	15,66	83,14	26,35	15,93	87,25	26,94	79,71	85,40	82,46	27,33	95,59	42,51	44,56
90	50	100	0,000001	0,5	15,98	83,77	26,84	15,26	87,60	25,99	79,07	85,38	82,10	27,12	95,39	42,24	44,29
91	50	100	0,000001	0,000001	15,69	83,15	26,40	15,22	87,11	25,91	79,46	85,49	82,36	27,86	95,29	43,12	44,45
92	50	100	0,000001	0,99999	87,48	5,84	10,95	88,44	1,34	2,64	87,18	1,05	2,08	94,51	3,88	7,46	5,78
93	50	100	0,99999	0,1	87,97	2,71	5,27	88,91	4,76	9,03	87,85	0,98	1,93	94,93	2,44	4,75	5,25
94	50	100	0,99999	0,5	87,47	5,60	10,53	88,90	4,57	8,69	87,22	5,99	11,21	94,11	3,68	7,08	9,38
95	50	100	0,99999	0,000001	87,34	5,90	11,06	88,47	6,63	12,33	87,25	3,15	6,07	94,53	1,87	3,67	8,28
96	50	100	0,99999	0,99999	87,19	6,63	12,32	88,74	6,15	11,51	87,40	4,01	7,67	94,48	5,31	10,05	10,39
97	500	1	0,1	0,1	87,69	83,15	85,36	88,34	87,70	88,02	79,94	77,37	78,63	94,43	86,74	90,42	85,61
98	500	1	0,1	0,5	87,28	83,85	85,53	88,81	87,96	88,38	79,82	77,73	78,76	94,74	86,93	90,67	85,83
99	500	1	0,1	0,000001	87,54	83,49	85,47	88,18	87,95	88,07	79,58	77,96	78,76	94,45	86,50	90,30	85,65
100	500	1	0,1	0,99999	87,12	63,17	73,24	88,69	67,80	76,85	78,41	68,57	73,16	94,85	86,77	90,63	78,47
101	500	1	0,9	0,1	87,04	83,50	85,23	88,56	87,90	88,23	78,00	68,08	72,70	94,66	86,39	90,34	84,13
102	500	1	0,9	0,5	87,21	83,24	85,18	88,38	87,14	87,75	78,12	68,32	72,89	94,72	86,26	90,29	84,03
103	500	1	0,9	0,000001	87,95	83,98	85,92	88,48	87,85	88,17	78,42	68,91	73,36	94,49	86,36	90,24	84,42
104	500	1	0,9	0,99999	87,50	63,75	73,76	88,43	67,97	76,86	78,67	68,87	73,45	94,77	86,97	90,70	78,69
105	500	1	0,000001	0,1	87,18	83,68	85,39	88,94	87,88	88,40	79,49	77,69	78,58	94,59	86,15	90,17	85,64
106	500	1	0,000001	0,5	87,11	83,51	85,27	88,80	87,07	87,93	79,45	77,31	78,37	94,60	86,96	90,62	85,55
107	500	1	0,000001	0,000001	87,17	83,55	85,32	88,10	87,60	87,85	79,31	77,88	78,59	94,77	86,63	90,52	85,57
108	500	1	0,000001	0,99999	87,74	63,93	73,97	88,90	67,10	76,48	78,29	68,03	72,80	94,50	86,79	90,48	78,43
109	500	1	0,99999	0,1	87,62	63,81	73,84	88,43	67,98	76,87	78,92	68,22	73,19	94,86	86,87	90,69	78,65
110	500	1	0,99999	0,5	87,24	63,57	73,55	88,22	67,34	76,38	78,82	68,05	73,04	94,35	86,33	90,16	78,28
111	500	1	0,99999	0,000001	87,63	63,43	73,59	88,62	67,07	76,35	78,35	68,37	73,02	94,46	86,68	90,40	78,34
112	500	1	0,99999	0,99999	87,41	63,31	73,44	88,69	67,65	76,76	78,76	68,07	73,03	94,05	86,60	90,17	78,35
113	500	50	0,1	0,1	23,11	63,66	33,91	17,32	67,55	27,57	87,49	85,81	86,64	57,63	86,27	69,10	54,31
114	500	50	0,1	0,5	23,19	63,56	33,98	17,34	67,05	27,56	87,83	85,00	86,39	57,82	86,21	69,22	54,29
115	500	50	0,1	0,000001	23,65	63,40	34,45	17,06	67,82	27,27	87,77	85,92	86,84	57,52	86,62	69,13	54,42
116	500	50	0,1	0,99999	87,76	4,39	8,37	88,44	1,89	3,70	87,09	2,43	4,72	94,27	4,52	8,62	6,35
117	500	50	0,9	0,1	87,42	4,72	8,96	88,33	3,12	6,03	87,15	68,69	76,83	61,77	68,54	64,98	39,20
118	500	50	0,9	0,5	87,77	1,06	2,09	88,64	4,64	8,81	87,01	68,52	76,67	61,91	68,90	65,22	38,20
119	500	50	0,9	0,000001	87,33	4,48	8,53	88,99	0,36	0,72	87,71	68,37	76,84	61,71	68,61	64,98	37,77
120	500	50	0,9	0,99999	87,89	6,63	12,33	88,76	3,67	7,05	87,08	3,69	7,07	94,70	6,78	12,65	9,78
121	500	50	0,000001	0,1	15,68	83,54	26,41	15,52	87,11	26,35	79,60	85,51	82,45	40,46	95,45	56,83	48,01
122	500	50	0,000001	0,5	15,94	83,02	26,75	15,85	87,24	26,83	79,67	85,91	82,67	40,31	95,95	56,77	48,26
123	500	50	0,000001	0,000001	15,79	83,43	26,56	15,68	87,15	26,57	79,12	85,35	82,12	40,22	95,53	56,61	47,96
124	500	50	0,000001	0,99999	87,50	2,39	4,65	88,05	6,58	12,25	87,84	0,83	1,65	0,24	2,66	0,44	4,75
125	500	50	0,99999	0,1	87,65	5,39	10,16	88,92	0,16	0,32	87,05	5,85	10,96	94,48	6,01	11,29	8,18
126	500	50	0,99999	0,5	87,93	4,85	9,19	88,18	4,57	8,69	87,86	6,48	12,06	94,14	2,10	4,11	8,51
127	500	50	0,99999	0,000001	87,11	2,66	5,16	88,54	5,49	10,34	87,54	2,23	4,36	94,59	6,29	11,79	7,91
128	500	50	0,99999	0,99999	87,40	1,69	3,32	88,52	5,83	10,94	87,89	1,38	2,71	94,51	5,62	10,62	6,90
129	500	100	0,1	0,1	12,32	83,56	21,48	20,64	67,82	31,65	79,57	85,20	82,29	49,59	95,14	65,19	50,15
130	500	100	0,1	0,5	12,72	83,48	22,08	20,20	67,98	31,15	79,40	85,93	82,54	49,48	95,78	65,25	50,25
131	500	100	0,1	0,000001	12,24	83,85	21,36	20,58	67,20	31,51	79,53	85,29	82,31	49,92	95,36	65,54	50,18
132	500	100	0,1	0,99999	87,28	2,86	5,55	88,64	6,97	12,93	87,13	5,27	9,94	94,84	5,13	9,73	9,53
133	500	100	0,9	0,1	87,79	5,63	10,59	88,22	1,94	3,79	87,25	1,53	3,01	94,94	6,45	12,08	7,37
134	500	100	0,9	0,5	87,15	3,01	5,81	88,93	3,30	6,36	87,48	4,77	9,04	94,92	4,23	8,10	7,33
135	500	100	0,9	0,000001	87,03	2,81	5,45	88,55	6,17	11,53	87,88	4,23	8,08	94,04	5,44	10,29	8,83
136	500	100	0,9	0,99999	87,05	2,51	4,88	88,59	0,90	1,77	87,54	5,39	10,15	94,45	0,92	1,83	4,66
137	500	100	0,000001	0,1	15,26	83,82	25,82	15,71	87,14	26,62	79,00	85,99	82,35	27,23	95,83	42,41	44,30
138	500	100	0,000001	0,5	15,67	83,92	26,41	15,23	87,75	25,96	79,53	85,74	82,52	27,31	95,28	42,45	44,34
139	500	100	0,000001	0,000001	15,81	83,23	26,57	15,70	87,20	26,61	79,15	85,23	82,08	27,91	95,86	43,23	44,62
140	500	100	0,000001	0,99999	87,98	1,42	2,80	88,73	1,57	3,08	87,94	1,73	3,39	94,66	3,79	7,29	4,14
141	500	100	0,99999	0,1	87,25	0,95	1,88	88,10	3,62	6,95	87,01	6,65	12,36	94,20	2,14	4,19	6,35
142	500	100	0,99999	0,5	87,93	3,97	7,60	88,39	4,60	8,74	87,92	5,78	10,84	94,81	4,41	8,43	8,90
143	500	100	0,99999	0,000001	87,41	2,05	4,01	88,12	1,73	3,39	87,77	1,30	2,56	94,02	4,80	9,13	4,77
144	500	100	0,99999	0,99999	87,23	6,93	12,84	88,26	4,71	8,94	87,46	1,70	3,34	94,05	5,74	10,82	8,99

D) Programa Batch com os principais comandos para executar o classificador Linear

PRIMEIRA EXECUÇÃO – C/ARQUIVO DE TREINAMENTO

```
java mkdict 2000 global.dit

java vectorize alimen ve.vec
java linear li1ctr.wts

java vectorize equipa ve.vec
java linear li2ctr.wts

java vectorize hext ve.vec
java linear li3ctr.wts

java vectorize honora ve.vec
java linear li4ctr.wts
```

SEGUNDA EXECUÇÃO – C/ARQUIVO DE TESTE

```
java vectorize alimen ve.vec
java testline li1ctr.wts li1ctepo.txt li1ctene.txt

java vectorize equipa ve.vec
java testline li2ctr.wts li2ctepo.txt li2ctene.txt

java vectorize hext ve.vec
java testline li3ctr.wts li3ctepo.txt li3ctene.txt

java vectorize honora ve.vec
java testline li4ctr.wts li4ctepo.txt li4ctene.txt
```


E) Resultados encontrados com o classificador Linear

							Alimentação			Equiparação			Horas Extras			Honorários			
Num reg	Num Pal	Freq min	Lim Dec	Lambda	Tx Apr	Tipo Car	prec	cob	med F	prec	cob	med F	prec	cob	med F	prec	cob	med F	med F media
1	30	1	0,2	0,01	0,25	binary	95,45	36,07	52,35	96,21	94,47	95,33	83,80	80,00	81,86	88,17	84,14	86,10	78,91
2	30	1	0,2	0,01	0,01	binary	95,16	0,89	1,76	96,66	74,83	84,36	92,48	79,36	85,42	88,22	84,22	86,17	64,43
3	30	1	0,2	0,01	0,01	tf*idf	95,30	36,31	52,58	96,31	54,04	69,23	92,26	88,63	90,41	88,54	84,55	86,50	74,68
4	30	1	0,2	0,9	0,25	binary	95,64	0,93	1,84	96,31	0,94	1,85	92,92	21,10	34,39	88,62	48,17	62,41	25,12
5	30	1	0,2	0,9	0,25	tf	95,83	6,07	11,41	96,95	54,51	69,78	92,48	71,74	80,80	88,53	84,26	86,34	62,09
6	30	1	0,2	0,9	0,25	tf*idf	95,73	1,81	3,54	96,90	74,35	84,14	92,80	88,98	90,85	88,31	84,94	86,59	66,28
7	30	1	0,2	0,9	0,9	binary	95,02	3,86	7,42	96,89	2,60	5,06	92,09	8,42	15,43	88,34	1,93	3,78	7,92
8	30	1	0,2	0,9	0,9	tf	95,55	1,07	2,12	96,75	54,02	69,33	92,50	63,20	75,09	88,91	84,47	86,63	58,29
9	30	1	0,2	0,9	0,9	tf*idf	95,94	5,24	9,94	96,35	74,01	83,72	92,83	63,03	75,08	88,95	84,46	86,65	63,85
10	30	1	0,2	0,9	0,01	binary	95,88	0,96	1,90	96,99	1,89	3,71	92,23	29,89	45,14	88,57	39,18	54,32	26,27
11	30	1	0,2	0,9	0,01	tf	95,89	3,41	6,58	96,16	54,31	69,42	92,06	79,45	85,29	88,48	84,31	86,35	61,91
12	30	1	0,2	0,9	0,01	tf*idf	95,85	2,05	4,02	96,38	54,53	69,65	92,92	88,71	90,77	88,71	84,04	86,31	62,69
13	30	1	0,2	0,00001	0,01	binary	95,69	2,47	4,82	96,63	74,89	84,38	92,93	79,34	85,60	88,71	84,89	86,76	65,39
14	30	1	0,2	0,00001	0,01	tf*idf	95,36	36,32	52,61	96,96	54,45	69,73	92,71	88,30	90,45	88,99	84,54	86,71	74,87
15	30	1	0,8	0,01	0,25	binary	95,15	20,76	34,09	96,22	74,44	83,94	92,26	63,52	75,24	88,46	66,02	75,61	67,22
16	30	1	0,8	0,01	0,25	tf	95,70	56,67	71,19	96,40	34,19	50,48	92,04	71,30	80,35	88,30	75,28	81,27	70,82
17	30	1	0,8	0,01	0,25	tf*idf	95,84	56,08	70,76	96,50	54,93	70,00	92,60	71,34	80,59	88,51	75,66	81,58	75,73
18	30	1	0,8	0,01	0,9	binary	95,27	20,87	34,24	96,90	74,25	84,08	92,00	54,58	68,52	88,30	66,87	76,10	65,73
19	30	1	0,8	0,01	0,9	tf	95,68	56,55	71,08	96,08	34,83	51,12	92,09	71,92	80,76	88,66	75,98	81,83	71,20
20	30	1	0,8	0,01	0,9	tf*idf	95,57	56,89	71,33	96,21	54,00	69,18	92,38	71,26	80,47	88,40	75,89	81,67	75,66
21	30	1	0,8	0,01	0,01	binary	95,96	4,34	8,31	96,60	4,61	8,81	92,18	0,99	1,97	88,26	78,41	83,04	25,53
22	30	1	0,8	0,01	0,01	tf	95,47	20,40	33,61	96,59	1,57	3,09	92,84	38,06	53,98	88,50	66,26	75,78	41,62
23	30	1	0,8	0,01	0,01	tf*idf	95,70	20,18	33,33	96,41	5,76	10,87	92,78	46,77	62,19	88,53	66,67	76,06	45,61
24	30	1	0,8	0,9	0,25	binary	95,30	4,97	9,45	96,37	3,89	7,48	92,06	3,91	7,50	88,40	6,26	11,69	9,03
25	30	1	0,8	0,9	0,25	tf	95,47	4,21	8,06	96,71	20,61	33,97	93,00	17,98	30,13	88,85	18,15	30,14	25,58
26	30	1	0,8	0,9	0,25	tf*idf	95,71	6,11	11,48	96,07	3,56	6,87	92,17	46,83	62,11	88,87	11,76	20,77	25,31
27	30	1	0,8	0,9	0,9	binary	95,30	1,73	3,40	96,23	3,96	7,60	92,86	5,10	9,67	88,61	3,99	7,63	7,08
28	30	1	0,8	0,9	0,9	tf	95,67	1,47	2,90	96,74	34,21	50,54	92,68	79,54	85,61	88,42	29,23	43,93	45,75
29	30	1	0,8	0,9	0,01	binary	95,17	1,99	3,90	96,58	2,95	5,73	92,03	1,40	2,75	88,06	2,04	3,99	4,09
30	30	1	0,8	0,9	0,01	tf	95,12	6,51	12,19	96,02	1,64	3,22	92,37	21,86	35,36	88,31	18,30	30,31	20,27
31	30	1	0,8	0,9	0,01	tf*idf	95,11	0,15	0,29	96,86	0,21	0,42	92,53	21,19	34,49	88,34	11,46	20,29	13,87
32	30	1	0,8	0,00001	0,25	binary	95,30	36,95	53,26	96,49	74,37	84,00	92,26	54,61	68,60	88,51	66,39	75,87	70,43
33	30	1	0,8	0,00001	0,9	binary	95,52	36,43	52,75	96,52	74,75	84,25	80,27	54,37	64,83	88,76	66,09	75,77	69,40
34	30	1	0,8	0,00001	0,01	binary	95,89	1,62	3,18	96,99	4,69	8,95	92,19	4,93	9,36	88,12	48,18	62,30	20,95
35	30	1	0,8	0,00001	0,01	tf	95,93	20,97	34,41	96,65	6,75	12,61	92,54	38,48	54,36	88,10	66,69	75,92	44,32
36	30	1	0,8	0,00001	0,01	tf*idf	95,22	20,03	33,10	96,86	5,40	10,22	92,25	46,66	61,97	88,05	66,64	75,86	45,29
37	30	50	0,2	0,01	0,01	binary	95,12	2,57	5,01	96,46	2,18	4,27	92,84	79,73	85,79	55,08	57,99	56,50	37,89
38	30	50	0,2	0,9	0,25	binary	95,19	2,54	4,95	96,77	2,28	4,45	92,13	29,39	44,57	88,39	4,13	7,88	15,46
39	30	50	0,2	0,9	0,25	tf	95,27	3,01	5,84	96,90	6,04	11,37	92,89	45,34	61,84	88,01	5,79	10,86	22,48
40	30	50	0,2	0,9	0,25	tf*idf	95,07	0,51	1,02	96,52	3,29	6,36	92,00	63,03	74,81	88,58	2,59	5,04	21,81
41	30	50	0,2	0,9	0,9	binary	95,25	5,51	10,41	96,53	1,31	2,59	92,05	8,11	14,91	88,63	4,56	8,68	9,15
42	30	50	0,2	0,9	0,9	tf	95,06	4,70	8,96	96,47	2,97	5,75	92,07	38,04	53,83	88,81	18,51	30,63	24,79
43	30	50	0,2	0,9	0,9	tf*idf	95,12	0,50	0,99	96,61	1,59	3,13	92,64	54,42	68,56	88,76	3,12	6,04	19,68
44	30	50	0,2	0,9	0,01	binary	95,63	0,79	1,58	96,04	4,45	8,51	92,85	21,70	35,18	88,55	5,82	10,93	14,05
45	30	50	0,2	0,9	0,01	tf	95,72	5,73	10,82	96,45	3,80	7,31	92,37	63,76	75,44	88,56	18,83	31,05	31,16
46	30	50	0,2	0,9	0,01	tf*idf	95,23	5,32	10,07	96,15	6,34	11,90	92,67	88,94	90,77	88,59	18,85	31,09	35,96
47	30	50	0,2	0,00001	0,01	binary	95,28	0,28	0,57	96,67	3,23	6,25	92,33	79,39	85,38	55,58	57,01	56,28	37,12
48	30	50	0,8	0,01	0,25	binary	95,99	2,11	4,13	96,64	4,04	7,76	92,34	54,87	68,84	88,36	11,64	20,58	25,33
49	30	50	0,8	0,01	0,01	binary	95,56	6,51	12,19	96,31	5,59	10,57	92,93	5,84	10,99	88,75	2,78	5,39	9,78
50	30	50	0,8	0,9	0,25	binary	95,08	0,44	0,87	96,17	1,34	2,64	92,31	1,35	2,66	88,61	0,25	0,50	1,67
51	30	50	0,8	0,9	0,25	tf	95,66	3,26	6,30	96,41	2,46	4,80	92,38	8,66	15,83	88,09	0,79	1,56	7,12
52	30	50	0,8	0,9	0,25	tf*idf	95,12	2,82	5,48	96,22	6,28	11,78	92,86	8,13	14,96	88,09	1,31	2,58	8,70
53	30	50	0,8	0,9	0,9	binary	95,54	1,69	3,32	96,05	5,34	10,12	92,68	18,00	30,14	88,80	2,46	4,78	12,09
54	30	50	0,8	0,9	0,9	tf	95,69	1,18	2,33	96,01	6,79	12,68	92,05	38,27	54,07	88,16	4,37	8,33	19,35
55	30	50	0,8	0,9	0,9	tf*idf	95,01	3,46	6,67	96,45	1,54	3,04	92,12	17,43	29,31	88,49	6,12	11,45	12,62
56	30	50	0,8	0,9	0,01	binary	95,14	4,74	9,02	96,25	0,65	1,30	92,33	1,68	3,31	88,23	2,38	4,64	4,57
57	30	50	0,8	0,9	0,01	tf	95,05	6,91	12,89	96,50	3,99	7,66	92,91	21,66	35,13	88,51	1,43	2,81	14,62
58	30	50	0,8	0,9	0,01	tf*idf	95,07	0,76	1,51	96,76	0,37	0,74	92,11	17,14	28,91	88,67	1,75	3,43	8,65
59	30	50	0,8	0,00001	0,25	binary	95,15	1,21	2,39	96,91	4,33	8,30	92,54	63,48	75,30	77,12	57,21	65,68	37,92
60	30	50	0,8	0,00001	0,9	binary	95,70	2,83	5,49	29,55	74,00	42,24	92,90	54,77	68,93	70,54	66,27	68,34	46,25
61	30	50	0,8	0,00001	0,01	binary	95,16	1,45	2,86	96,15	3,57	6,88	92,94	8,72	15,94	88,46	18,31	30,34	14,01
62	50	1	0,2	0,01	0,25	binary	95,71	36,83	53,20	96,09	94,47	95,27	84,35	96,25	89,91	88,95	84,06	86,44	81,20
63	50	1	0,2	0,01	0,9	binary	95,43	36,94	53,26	96,96	94,58	95,76	84,37	96,95	90,22	88,93	84,39	86,60	81,46
64	50	1	0,2	0,01	0,01	binary	95,99	6,46	12,10	96,43	34,87	51,22	92,20	71,16	80,32	88,88	84,18	86,47	57,53
65	50	1	0,2	0,9	0,25	binary	95,89	4,63	8,84	96,48	3,69	7,10	92,82	71,74	80,93	88,60	66,63	76,06	43,23
66	50	1	0,2	0,9	0,9	binary	95,26	6,10	11,47	96,63	20,73	34,13	92,32	63,61	75,32	88,85	0,05	0,10	30,26
67	50	1	0,2	0,9	0,01	binary	95,36	0,83	1,65	96,09	4,13	7,93	92,27	46,82	62,12	88,77	57,88	70,07	35,44
68	50	1	0,2	0,00001	0,01	binary	95,73	1,03	2,03	96,02	34,67	50,94	92,83	72,00	81,10	88,43	84,80	86,58	55,16
69	50	1	0,2	0,00001	0,01	tf	95,69	36,27	52,60	96,65	74,63	84,23	92,82	88,81	90,77	88,96	84,49	86,67	78,57
70	50	1	0,2	0,00001	0,01	tf*idf	95,58	36,40	52,72	96,75	74,53	84,20	92,75	88,24	90,44	88,49	84,80	86,61	78,49
71	50	1	0,8	0,01	0,25	binary	95,54	2,24	4,38	96,59	34,89	51,26	92,90	46,45	61,83	88,22	66,64	75,93	48,35
72	50	1	0,8	0,01	0,9	binary	95,87	6,01	11,31	96,34	34,57	50,88	92,45	46,00	61,44	88,17	66,03	75,51	49,78
73	50	1	0,8	0,01	0,01	binary	95,50	4,75	9,04	96,13	3,53	6,81	92,39	4,08	7,81	88,56	48,29	62,50	21,54
74	50	1	0,8	0,9	0,25	binary	95,08	1,58	3,10	96,56	5,05	9,59	92,41	2,96	5,74	88,97	3,98	7,63	6,52
75	50	1	0,8	0,9	0,9	binary	95,87	1,63	3,21	96,91	3,60	6,95	92,43	4,38	8,37	88,39	11,21	19,90	9,61
76	50	1	0,8	0,00001	0,9	binary	95,84	4,94	9,40	96,93	34,10	50,45	92,27	54,73	68,70	88,68	66,33	75,89	51,11
77	50	1	0,8	0,00001	0,01	binary	95,92	5,70	10,77	96,74	2,36	4,61	92,88	6,79	12,65	88,46	48,77	62,87	22,73
78	50	50	0,2	0,01	0,25	binary	95,64	20,07	33,18	96,98	20,90	34,39	92,55	88,14	90,29	61,69	57,50	59,52	54,35
79	50	50	0,2	0,01	0,9	binary	95,97	20,94	34,38	96,60	34,67	51,03	92,10	88,79	90,42	68,76	57,54	62,65	59,62
80	50	50	0,2	0,01	0,01	binary	95,71	4,04	7,76	96,80	2,25	4,40	92,23	79,44	85,36	61,89	57,14	59,42	39,24
81	50	50	0,2	0,9	0,25	binary	95,37	6,20	11,64	96,24	1,13	2,24	92,86	3,76	7,23	88,27	4,39	8,36	7,37
82	50	50	0,2	0,9	0,9	binary	95,29	1,43	2,81	96,94	3,47	6,69	92,96	71,21	80,64	88,01	1,43	2,81	23,24
83	50	50	0,2	0,9	0,01	binary	95,78	4,12	7,90	96,94	3,64	7,01	92,57	21,79	35,27	88,43	0,22	0,44	12,65
84	50	50	0,2	0,00001	0,01	binary	95,39	4,33	8,28	96,54	0,34	0,68	92,91	79,12	85,46	55,20	57,44	56,30	37,68
85	50	50	0,8	0,01	0,25	binary	95,31	4,38	8,38	96,01	5,85	11,04	92,76	54,13	68,37	88,92	39,56	54,76	35,64
86	50	50	0,8	0,01	0,01	binary	95,76	5,90	11,12	96,60	0,95	1,88	92,27	3,88	7,45	88,49	4,14	7,91	7,09
87	50	50	0,8	0,9	0,25	binary	95,70	4,45	8,51	96,90	0,45	0,89	92,38	5,40	10,21	88,55	6,10	11,42	7,75
88	50	50	0,8	0,9	0,9	binary	95,71	4,15	7,96	96,56	2,04	4,00	92,37	5,28	10,00	88,80	5,63	10,59	8,14
89	50	50	0,8	0,9	0,9	tf	95,65	4,83	9,20	96,58	0,23	0,46	92,28	21,10	34,34	88,39	5,94	11,12	13,78
90	50	50	0,8	0,9	0,9	tf*idf	95,35	6,52	12,21	96,45	0,08	0,17	92,87	71,94	81,07	88,82	5,22	9,86	25,83
91	50	50	0,8	0,9	0,01	binary	95,97	3,64	7,01	96,32	4,88	9,29	92,68	6,56	12,26	88,05	2,80	5,42	8,49
92	50	50	0,8	0,00001	0,25	binary	95,89	20,18	33,34	96,55	4,26	8,16	92,63	63,45	75,31	88,79	39,77	54,94	42,94
93	50	50	0,8	0,00001	0,9	binary	95,03	20,75	34,06	96,76	3,32	6,41	92,66	63,66	75,47	71,27	29,04	41,27	39,30
94	50	50	0,8	0,00001	0,01	binary	95,72	4,63	8,84	96,71	0,88	1,75	92,67	8,74	15,98	88,46	18,18	30,16	14,18
95	500	1	0,2	0,01	0,25	binary	95,55	56,72	71,19	96,22	74,47	83,96	92,08	79,56	85,36	88,24	84,53	86,34	81,71
96	500	1	0,2	0,01	0,9	binary	95,08	56,89	71,19	96,86	74,44	84,19	92,60	79,59	85,61	88,91	84,86	86,84	81,95
97	500	1	0,2	0,01	0,01	binary	95,08	3,35	6,48	96,51	0,17	0,33	92,63	21,59	35,01	88,20	84,04	86,07	31,97
98	500	1	0,2	0,9	0,25	binary	95,14	0,59	1,16	96,10	0,20	0,40	92,40	54,03	68,19	88,54	0,73	1,44	17,80
99	500	1	0,2	0,9	0,9	binary	95,94	1,36	2,68	96,74	1,76	3,46	92,33	38,25	54,10	88,08	75,15	81,10	35,33
100	500	1	0,2	0,9	0,01	binary	95,38	0,47	0,94	96,12	2,26	4,41	92,41	29,42	44,63	88,09	84,41	86,21	34,05
101	500	1	0,2	0,00001	0,25	binary	95,01	56,03	70,49	96,09	74,48	83,92	92,20	79,18	85,20	88,02	84,22	86,08	81,42

102	500	1	0,8	0,01	0,25	binary	95,65	1,71	3,35	96,97	1,72	3,38	92,99	1,46	2,87	88,81	75,97	81,89	22,87
103	500	1	0,8	0,01	0,9	binary	95,75	5,84	11,00	96,25	5,07	9,63	92,57	6,04	11,34	88,63	75,33	81,44	28,35
104	500	1	0,8	0,01	0,01	binary	95,10	1,04	2,05	96,77	2,23	4,36	92,93	1,83	3,59	88,17	29,89	44,65	13,66
105	500	1	0,8	0,00001	0,25	binary	95,47	20,49	33,74	96,41	2,62	5,10	92,40	5,11	9,68	88,52	75,85	81,70	32,56
106	500	1	0,8	0,00001	0,9	binary	95,57	20,70	34,04	96,37	4,42	8,46	92,38	4,16	7,96	88,72	75,18	81,39	32,96
107	500	1	0,8	0,00001	0,01	binary	95,48	4,67	8,91	96,31	5,01	9,52	92,71	5,36	10,13	88,46	39,06	54,19	20,69
108	30	1	0,2	0,01	0,25	tf	78,21	96,14	86,25	96,62	94,73	95,67	92,81	88,71	90,72	88,66	84,09	86,31	89,74
109	500	1	0,2	0,01	0,25	tf*idf	78,96	96,00	86,65	96,90	94,77	95,82	92,83	88,82	90,78	88,63	84,14	86,33	89,90
110	500	1	0,2	0,01	0,9	tf	78,95	96,57	86,88	96,18	94,71	95,44	92,93	88,96	90,90	88,88	84,55	86,66	89,97
111	500	1	0,2	0,01	0,9	tf*idf	78,14	96,61	86,40	96,57	94,42	95,48	92,12	88,80	90,43	88,98	84,97	86,92	89,81
112	30	1	0,2	0,00001	0,25	tf*idf	78,82	96,34	86,70	97,00	94,26	95,61	92,40	88,01	90,15	88,25	84,47	86,32	89,70
113	30	1	0,2	0,00001	0,9	tf*idf	78,46	96,44	86,53	96,90	94,67	95,78	92,42	88,97	90,67	88,20	84,49	86,31	89,82
114	30	1	0,8	0,00001	0,25	tf	75,75	76,94	76,34	96,16	34,02	50,26	92,33	79,14	85,23	88,66	75,85	81,75	73,40
115	30	1	0,8	0,00001	0,25	tf*idf	75,61	76,23	75,92	96,84	54,30	69,58	93,00	71,33	80,73	88,74	75,29	81,47	76,93
116	30	1	0,8	0,00001	0,9	tf	75,56	76,15	75,85	96,04	34,41	50,67	92,72	79,44	85,57	88,95	75,99	81,96	73,51
117	30	1	0,2	0,00001	0,25	binary	70,96	56,44	62,87	96,47	94,37	95,41	84,04	88,56	86,24	88,26	75,91	81,62	81,54
118	30	1	0,8	0,00001	0,9	tf*idf	70,26	56,27	62,49	96,53	54,81	69,92	92,90	71,02	80,50	88,59	75,52	81,54	73,61
119	30	1	0,2	0,00001	0,25	tf	66,73	96,72	78,97	96,22	94,66	95,44	92,72	79,07	85,35	88,69	84,84	86,72	86,62
120	30	1	0,2	0,00001	0,9	tf	66,72	96,55	78,91	96,33	94,23	95,27	92,43	88,06	90,19	88,82	84,24	86,47	87,71
121	30	1	0,2	0,01	0,9	binary	62,75	36,79	46,39	96,45	94,52	95,47	84,70	88,40	86,51	88,12	84,50	86,27	78,66
122	30	1	0,2	0,01	0,01	tf	62,57	36,75	46,30	96,02	54,95	69,90	92,34	79,73	85,58	88,78	84,35	86,51	72,07
123	30	1	0,2	0,00001	0,9	binary	62,55	36,87	46,39	96,64	94,38	95,49	84,30	88,84	86,51	79,15	75,49	77,28	76,42
124	30	1	0,2	0,00001	0,01	tf	62,39	36,52	46,07	96,84	54,40	69,66	92,05	79,00	85,03	88,91	84,71	86,76	71,88
125	30	50	-0,8	0,00001	0,9	tf	58,87	96,37	73,09	23,74	74,96	36,06	92,99	88,74	90,82	65,93	75,20	70,26	67,56
126	50	1	-0,8	0,01	0,25	binary	58,02	96,57	72,49	67,76	94,13	78,80	59,51	96,70	73,68	57,29	84,90	68,42	73,35
127	50	1	-0,8	0,01	0,25	tf*idf	58,26	96,35	72,61	59,35	94,80	73,00	67,69	96,73	79,65	49,78	84,28	62,59	71,96
128	50	1	-0,8	0,01	0,9	binary	58,71	96,49	73,00	67,52	94,82	78,88	59,02	96,52	73,25	57,44	84,95	68,54	73,42
129	50	1	-0,8	0,00001	0,25	binary	58,94	96,64	73,22	67,58	94,67	78,86	72,26	96,25	82,55	57,84	84,54	68,69	75,83
130	30	1	-0,8	0,00001	0,25	binary	51,27	96,07	66,86	79,62	94,81	86,55	78,58	96,38	86,57	57,62	84,48	68,51	77,12
131	30	1	-0,8	0,00001	0,9	binary	51,52	96,33	67,13	67,76	94,92	79,07	72,72	96,56	82,96	61,98	84,82	71,62	75,20
132	50	1	-0,8	0,01	0,25	tf	51,79	96,55	67,42	52,97	94,30	67,83	77,84	88,17	82,69	53,81	84,94	65,88	70,96
133	50	1	-0,8	0,00001	0,9	binary	51,21	96,38	66,89	67,88	94,87	79,14	67,03	96,75	79,19	57,51	84,27	68,37	73,40
134	50	50	-0,8	0,00001	0,9	binary	51,81	96,75	67,49	29,75	20,77	24,46	71,91	88,95	79,53	57,45	66,49	61,64	58,28
135	30	1	-0,8	0,01	0,9	binary	50,67	96,84	66,53	67,32	94,08	78,48	63,28	96,34	76,39	57,04	84,61	68,14	72,39
136	30	1	-0,8	0,01	0,25	binary	45,04	96,15	61,34	67,18	94,52	78,54	59,37	96,46	73,50	57,98	84,15	68,66	70,51
137	30	1	-0,8	0,01	0,25	tf*idf	45,74	96,29	62,02	79,33	94,79	86,37	72,13	96,02	82,37	80,95	84,71	82,79	78,39
138	30	1	-0,8	0,01	0,9	tf*idf	45,04	96,99	61,51	79,93	94,42	86,57	72,81	96,44	82,97	80,47	84,72	82,54	78,40
139	30	1	-0,8	0,00001	0,9	tf*idf	45,28	96,20	61,58	79,71	94,53	86,49	78,40	96,99	86,71	80,12	84,29	82,15	79,23
140	30	50	-0,8	0,01	0,9	tf	45,62	96,48	61,95	24,05	94,85	38,38	78,00	88,61	82,97	53,88	84,65	65,85	62,28
141	30	50	0,2	0,01	0,25	tf*idf	45,11	36,47	40,33	71,52	54,81	62,06	92,70	88,55	90,57	59,90	75,98	66,99	64,99
142	30	50	0,2	0,01	0,9	binary	45,63	20,51	28,30	96,57	20,33	33,59	92,19	88,58	90,35	68,42	57,50	62,49	53,68
143	30	50	0,2	0,01	0,9	tf	45,47	36,68	40,61	96,33	54,86	69,91	92,12	79,30	85,23	55,70	75,19	63,99	64,93
144	30	50	0,2	0,01	0,9	tf*idf	45,45	36,79	40,66	96,56	54,33	69,54	92,82	88,34	90,53	59,24	75,82	66,51	66,81
145	30	50	0,2	0,00001	0,25	tf	45,98	56,82	50,83	96,23	34,70	51,01	92,85	79,61	85,72	67,77	84,00	75,02	65,65
146	30	50	0,2	0,00001	0,25	tf*idf	45,05	36,21	40,15	71,43	54,50	61,83	92,50	88,58	90,50	73,41	84,25	78,46	67,73
147	30	50	0,2	0,00001	0,9	binary	45,54	20,03	27,82	96,06	20,43	33,69	92,19	88,81	90,47	57,47	66,56	61,68	53,41
148	30	50	0,8	0,01	0,25	tf*idf	45,99	20,25	28,11	96,29	20,12	33,29	92,49	71,81	80,85	68,91	20,13	31,15	43,35
149	30	50	0,8	0,01	0,9	tf	45,50	20,09	27,88	96,05	34,33	50,58	92,65	54,37	68,52	68,63	20,58	31,66	44,66
150	30	50	0,8	0,01	0,9	tf*idf	45,80	36,14	40,40	96,76	20,31	33,57	92,37	71,52	80,62	63,05	11,36	19,24	43,46
151	30	50	0,8	0,00001	0,9	tf	45,38	56,42	50,30	53,88	74,78	62,64	92,57	54,91	68,93	59,92	75,03	66,63	62,12
152	50	50	0,2	0,00001	0,9	binary	45,25	36,78	40,58	96,56	0,19	0,37	92,74	71,50	80,75	65,37	75,51	70,07	47,94
153	30	1	-0,8	0,00001	0,25	tf*idf	40,53	96,91	57,16	79,56	94,35	86,33	84,92	96,35	90,27	80,99	84,96	82,93	79,17
154	30	1	-0,8	0,00001	0,9	tf	40,37	96,59	56,94	67,56	94,52	78,80	84,41	88,77	86,53	80,22	84,61	82,36	76,16
155	30	50	-0,8	0,01	0,9	binary	39,25	76,61	51,91	29,71	94,75	45,23	71,66	88,77	79,30	40,66	84,01	54,80	57,81
156	30	1	-0,8	0,01	0,9	tf	37,24	96,19	53,69	67,33	94,85	78,75	84,72	96,12	90,06	80,72	84,38	82,51	76,25
157	30	1	-0,8	0,00001	0,25	tf	37,06	96,37	53,54	67,81	94,69	79,03	92,06	96,45	94,20	80,95	84,64	82,75	77,38
158	30	50	-0,8	0,00001	0,25	tf	37,60	96,62	54,13	34,98	54,73	42,68	92,83	88,36	90,54	53,35	84,71	65,47	63,21
159	30	50	-0,8	0,00001	0,9	tf*idf	35,43	76,55	48,44	26,68	54,83	35,90	92,51	88,14	90,27	59,68	75,75	66,76	60,34
160	30	50	0,2	0,01	0,25	tf	35,56	36,13	35,84	71,35	54,20	61,60	92,92	79,96	85,95	65,96	75,11	70,24	63,41
161	30	50	0,2	0,00001	0,9	tf*idf	35,25	36,57	35,89	36,19	74,08	48,63	92,89	79,67	85,77	57,31	84,13	68,18	59,62
162	50	50	-0,8	0,00001	0,25	binary	35,20	76,65	48,24	17,13	94,21	28,98	65,23	88,89	75,24	57,07	84,89	68,26	55,18
163	30	1	-0,8	0,01	0,25	tf	33,30	96,74	49,54	67,40	94,31	78,61	84,67	96,80	90,33	80,38	84,58	82,43	75,23
245	30	50	-0,8	0,00001	0,25	binary	33,85	96,23	50,08	26,33	54,55	35,52	65,57	88,29	75,25	55,39	75,56	63,92	56,19
164	30	100	-0,8	0,01	0,25	binary	20,49	96,47	33,80	15,07	94,70	26,00	84,17	96,99	90,13	43,17	10,57	16,98	41,72
165	30	100	-0,8	0,01	0,25	tf	19,85	96,93	32,96	17,14	94,11	28,99	92,26	88,06	90,11	41,88	75,73	53,93	51,50
166	30	100	-0,8	0,01	0,25	tf*idf	16,53	96,64	28,24	17,95	94,16	30,15	92,33	88,36	90,30	41,55	75,82	53,68	50,59
167	30	100	-0,8	0,01	0,9	binary	0,50	0,65	0,57	5,36	0,36	0,68	6,15	0,43	0,80	2,15	1,30	1,62	0,92
168	30	100	-0,8	0,01	0,9	tf	21,89	96,54	35,69	19,97	94,58	32,98	92,37	88,63	90,46	41,24	75,93	53,45	53,15
169	30	100	-0,8	0,01	0,9	tf*idf	2,13	2,68	2,38	5,66	3,44	4,28	3,77	2,21	2,79	1,75	5,71	2,68	3,03
170	30	100	-0,8	0,01	0,01	binary	3,98	5,11	4,47	4,52	0,30	0,57	0,83	2,04	1,18	5,80	6,27	6,02	3,06
171	30	100	-0,8	0,01	0,01	tf	5,58	6,87	6,16	3,30	6,75	4,44	4,30	6,30	5,11	4,39	5,19	4,76	5,12
172	30	100	-0,8	0,01	0,01	tf*idf	0,07	6,68	0,13	1,74	3,61	2,35	0,21	6,05	0,40	3,58	1,20	1,80	1,17
173	30	100	-0,8	0,9	0,25	binary	6,24	6,82	6,52	6,86	6,21	6,52	6,70	6,09	6,38	2,72	0,23	0,43	4,96
174	30	100	-0,8	0,9	0,25	tf	2,66	4,11	3,23	3,43	0,06	0,11	5,92	4,73	5,26	5,42	6,62	5,96	3,64
175	30	100	-0,8	0,9	0,25	tf*idf	4,12	2,91	3,41	2,46	6,31	3,55	4,25	4,22	4,23	0,90	6,96	1,59	3,19
176	30	100	-0,8	0,9	0,9	binary	18,23	96,05	30,65	23,74	54,22	33,02	28,42	96,93	43,96	38,27	84,18	52,61	40,06
177	30	100	-0,8	0,9	0,9	tf	0,56	0,60	0,58	6,62	4,52	5,37	3,95	6,37	4,88	5,02	4,73	4,87	3,92
178	30	100	-0,8	0,9	0,9	tf*idf	4,58	2,93	3,57	5,49	4,38	4,87	0,53	3,22	0,91	4,83	1,73	2,55	2,98
179	30	100	-0,8	0,9	0,01	binary	1,03	2,39	1,44	2,45	3,56	2,90	1,71	0,22	0,39	3,54	3,50	3,52	2,06
180	30	100	-0,8	0,9	0,01	tf	3,68	3,36	3,51	2,49	1,50	1,88	1,60	1,68	1,64	2,07	2,27	2,16	2,30
181	30	100	-0,8	0,9	0,01	tf*idf	6,42	4,54	5,32	3,41	3,05	3,22	3,13	2,54	2,80	3,25	4,10	3,63	3,74
182	30	100	-0,8	0,00001	0,25	binary	20,03	96,97	33,20	16,97	94,30	28,77	84,93	96,23	90,23	38,56	84,27	52,91	51,28
183	30	100	-0,8	0,00001	0,25	tf	4,77	1,16	1,87	6,66	1,15	1,96	3,04	2,28	2,60	5,12	5,15	5,14	2,89
184	30	100	-0,8	0,00001	0,25	tf*idf	0,69	3,28	1,15	3,23	4,87	3,88	0,20	5,53	0,39	0,37	1,36	0,58	1,50
185	30	100	-0,8	0,00001	0,9	binary	6,26	4,58	5,29	0,80	1,43	1,02	1,60	0,28	0,47	6,45	4,02	4,95	2,94
186	30	100	-0,8	0,00001	0,9	tf	0,73	6,04	1,31	2,54	6,30	3,62	0,45	6,09	0,84	0,55	1,41	0,79	1,64
187	30	100	-0,8	0,00001	0,9	tf*idf	3,05	6,70	4,19	5,75	1,69	2,61	5,24	1,60	2,45	2,35	5,80	3,35	3,15
188	30	100	-0,8	0,00001	0,01	binary	2,40	3,92	2,98	4,51	3,83	4,14	4,16	5,60	4,77	5,87	6,12	6,00	4,47
189	30	100	-0,8	0,00001	0,01	tf	5,91	5,83	5,87	3,69	5,50	4,42	5,55	5,51	5,53	1,14	2,53	1,57	4,35
190	30	100	-0,8	0,00001	0,01	tf*idf	6,39	3,39	4,43	0,38	3,12	0,68	2,35	1,17	1,56	1,83	1,66	1,74	2,10
191	30	100	0,2	0,01	0,25	binary	95,97	0,42	0,83	96,68	5,04	9,58	85,00	96,22	90,26	32,24	20,33	24,93	31,40
192	30	100	0,2	0,01	0,25	tf	4,24	5,83	4,91	6,02	6,58	6,29	3,00	1,71	2,18	4,16	5,28	4,66	4,51
193	30	100	0,2	0,01	0,25	tf*idf	0,18	1,49	0,33	5,79	2,48	3,47	0,94	3,75	1,51	1,95	0,88	1,21	1,63
194	30	100	0,2	0,01	0,9	binary	95,33	3,19	6,18	96,56	2,74	5,33	84,84	96,54	90,32	88,05	1,98	3,88	26,43
195	30	100	0,2	0,01	0,9	tf	1,37	1,74	1,53	2,84	1,20	1,69	5,12	5,31	5,21	5,99	4,68	5,26	3,42
196	30	100	0,2	0,01	0,9	tf*idf	3,82	4,51	4,13	5,97	5,05	5,47	6,98	3,03	4,23	4,61	1,93	2,72	4,14
197	30	100	0,2	0,01	0,01	binary	3,66	2,95	3,27	1,69	3,87	2,36	1,45	1,79	1,60	3,59	2,82	3,16	2,60
198	30	100	0,2	0,01	0,01	tf	1,71	1,75	1,73	5,73	2,86	3,82	5,22	1,77	2,64	5,69	1,16	1,93	2,53
199	30	100	0,2	0,01	0,01	tf*idf	0,38	1,68	0,62	1,51	5,66	2,39	4,68	6,61	5,48	4,55	3,21	3,76	3,06
200	30	100	0,2	0,9	0,25	binary	4,81	1,49	2,28	5,30	5,19	5,24	0,50	4,06	0,90	3,85	3,46	3,65	3,02
201	30	100	0,2	0,9	0,25	tf	2,87	0,75	1,18	2,33	0,72	1,10	0,62	6,18	1,13	5,32	3,56	4,26	1,92
202	30	100	0,2	0,9	0,25	tf*idf	1,32	0,29	0,48	0,53	4,21	0,95	0,20	5,24	0,38	2,54	5,61	3,50	1,33
203	30	100	0,2	0,9	0,9	binary	3,88	5,45	4,53	2,26	2,91	2,55	3,41	4,83	4,00	0,83	6,03	1,46	3,13
204	30	100	0,2	0,9	0,9	tf	2,10	4,30	2,82	0,49	6,19	0,92	1,74	4,33	2,49	4,29	3,34	3,75	2,49
205	30	100	0,2	0,9	0,9	tf*idf	4,50	1,33	2,05	2,36	2,11	2,23	5,74	6,51	6,10	6,84	5,95	6,37	4,19
206	30	100	0,2	0,9	0,01	binary	6,57	0,78	1,40	0,37	3,12	0,66	5,93	2,56	3,58	4,07	5,36	4,63	2,57
207	30	100	0,2	0,9	0,01	tf	1,50	6,83	2,46	4,94	1,17	1,89	6,71	4,91	5,67	6,81	2,83	4,00	3,50
208	30	100	0,2	0,9	0,01	tf*idf	0,26	0,86	0,41	1,67	5,94	2,60	6,35	4,44	5,23	2,60	4,19	3,21	2,86
209	30	100	0,2	0,00001	0,25	binary	2,42	4,14	3,05	0,11	6,60	0,21	1,65	5,01	2,48	4,59	2,24	3,01	2,19
210	30	100	0,2	0,00001	0,25	tf	3,53	3,93	3,72	3,63	2,23	2,76	1,06	2,51	1,49	1,03	6,40	1,78	2,44
211	30	100	0,2	0,00001	0,25	tf*idf	3,51	3,36	3,43	4,46	1,64	2,39	5,79	0,09	0,19	0,61	3,08	1,02	1,76
212	30	100	0,2	0,00001	0,9	binary	20,23	96,21	33,42	96,80	4,77	9,08	84,65	96,30	90,10	32,53	20,28	24,98	39,40

213	30	100	0,2	0,00001	0,9	tf	2,35	0,39	0,66	6,95	2,58	3,76	5,66	2,23	3,20	5,93	1,52	2,42	2,51
214	30	100	0,2	0,00001	0,9	tf*idf	1,13	6,57	1,92	2,05	3,78	2,66	5,79	5,52	5,65	2,79	1,94	2,29	3,13
215	30	100	0,2	0,00001	0,01	binary	95,49	5,41	10,24	96,53	2,91	5,65	84,40	96,29	89,95	32,90	20,90	25,56	32,85
216	30	100	0,2	0,00001	0,01	tf	5,71	6,44	6,05	2,59	5,23	3,46	0,15	1,94	0,28	5,18	6,51	5,77	3,89
217	30	100	0,2	0,00001	0,01	tf*idf	4,68	6,76	5,53	1,21	6,30	2,03	1,65	3,57	2,26	4,10	0,76	1,28	2,78
218	30	100	0,8	0,01	0,25	binary	95,76	4,05	7,77	96,86	2,55	4,98	92,02	0,87	1,72	88,08	3,74	7,17	5,41
219	30	100	0,8	0,01	0,25	tf	1,08	1,23	1,15	4,99	5,89	5,40	0,99	6,77	1,72	5,63	5,59	5,61	3,47
220	30	100	0,8	0,01	0,25	tf*idf	4,66	1,17	1,88	1,84	4,15	2,55	0,07	2,17	0,14	3,75	5,37	4,42	2,25
221	30	100	0,8	0,01	0,9	binary	1,89	4,21	2,61	0,82	3,16	1,30	6,72	6,62	6,67	4,31	0,29	0,54	2,78
222	30	100	0,8	0,01	0,9	tf	95,28	0,25	0,50	6,02	2,17	3,18	92,53	79,74	85,66	88,57	4,78	9,08	24,61
223	30	100	0,8	0,01	0,9	tf*idf	4,12	4,76	4,42	2,08	0,86	1,22	6,58	2,42	3,54	0,30	2,56	0,54	2,43
224	30	100	0,8	0,01	0,01	binary	0,54	1,40	0,78	0,48	3,76	0,86	0,25	3,66	0,47	5,14	6,05	5,55	1,92
225	30	100	0,8	0,01	0,01	tf	2,56	2,91	2,72	6,82	1,46	2,41	6,16	2,04	3,07	4,21	4,57	4,38	3,14
226	30	100	0,8	0,01	0,01	tf*idf	3,71	1,13	1,73	2,05	5,16	2,94	0,24	4,96	0,45	0,68	0,70	0,69	1,45
227	30	100	0,8	0,9	0,25	binary	1,49	2,52	1,87	5,80	5,54	5,67	2,41	6,62	3,53	3,51	2,29	2,77	3,46
228	30	100	0,8	0,9	0,25	tf	5,10	0,04	0,07	4,19	2,96	3,47	0,31	1,84	0,53	3,39	4,04	3,69	1,94
229	30	100	0,8	0,9	0,25	tf*idf	2,51	3,86	3,04	3,17	6,32	4,22	5,87	3,05	4,02	5,60	5,79	5,69	4,24
230	30	100	0,8	0,9	0,9	binary	3,55	0,69	1,15	1,55	3,43	2,14	5,53	1,30	2,11	4,00	0,86	1,42	1,70
231	30	100	0,8	0,9	0,9	tf	6,85	0,69	1,26	1,46	5,62	2,32	4,36	3,81	4,07	4,66	0,04	0,08	1,93
232	30	100	0,8	0,9	0,9	tf*idf	0,02	6,88	0,03	6,15	1,00	1,72	1,10	6,40	1,87	0,18	1,73	0,33	0,99
233	30	100	0,8	0,9	0,01	binary	6,63	5,79	6,18	7,00	5,55	6,19	2,85	3,87	3,28	5,02	6,45	5,64	5,32
234	30	100	0,8	0,9	0,01	tf	2,19	4,93	3,03	0,89	1,87	1,20	2,81	1,28	1,75	3,96	6,29	4,86	2,71
235	30	100	0,8	0,9	0,01	tf*idf	0,44	0,58	0,50	3,51	0,48	0,84	6,50	2,32	3,42	0,83	2,59	1,25	1,51
236	30	100	0,8	0,00001	0,25	binary	0,52	1,03	0,69	4,78	0,92	1,55	3,65	3,64	3,65	3,17	6,42	4,25	2,53
237	30	100	0,8	0,00001	0,25	tf	2,31	6,35	3,39	1,20	1,07	1,13	1,05	6,94	1,83	2,29	2,54	2,41	2,19
238	30	100	0,8	0,00001	0,25	tf*idf	4,70	3,04	3,69	2,87	2,68	2,77	6,38	3,41	4,45	1,25	5,00	2,00	3,23
239	30	100	0,8	0,00001	0,9	binary	1,50	2,02	1,73	4,17	1,35	2,04	5,34	4,42	4,84	4,72	0,68	1,18	2,45
240	30	100	0,8	0,00001	0,9	tf	3,65	2,45	2,93	3,64	1,85	2,46	5,11	6,25	5,63	1,20	1,67	1,40	3,10
241	30	100	0,8	0,00001	0,9	tf*idf	95,56	3,94	7,57	96,48	0,74	1,48	92,81	88,69	90,71	88,71	3,45	6,65	26,60
242	30	100	0,8	0,00001	0,01	binary	4,13	1,52	2,22	2,37	3,11	2,69	6,67	6,67	6,67	5,29	3,31	4,07	3,91
243	30	100	0,8	0,00001	0,01	tf	2,22	5,84	3,21	5,33	3,84	4,46	4,63	3,45	3,95	5,55	6,75	6,09	4,43
244	30	100	0,8	0,00001	0,01	tf*idf	6,87	3,08	4,26	3,55	6,15	4,50	4,84	3,68	4,18	2,93	5,24	3,75	4,17
246	30	50	-0,8	0,00001	0,01	tf*idf	33,28	96,15	49,44	20,47	94,66	33,67	84,72	96,16	90,08	40,64	84,59	54,90	57,02
247	500	1	-0,8	0,01	0,25	binary	33,30	96,22	49,47	34,84	94,17	50,86	52,39	96,77	67,98	53,93	84,16	65,74	58,51
248	500	1	-0,8	0,01	0,9	binary	33,83	96,94	50,15	34,44	94,27	50,45	52,35	96,98	67,99	53,17	84,68	65,32	58,48
249	500	1	-0,8	0,00001	0,25	binary	33,88	96,81	50,20	34,24	94,71	50,30	59,36	96,17	73,41	53,28	84,16	65,25	59,79
250	500	1	-0,8	0,00001	0,9	binary	33,81	96,47	50,07	34,39	94,77	50,47	59,49	96,01	73,46	53,28	84,49	65,35	59,84
251	30	50	-0,8	0,00001	0,9	binary	31,40	76,18	44,47	23,59	54,92	33,00	71,58	88,83	79,28	40,66	84,08	54,82	52,89
252	30	50	-0,8	0,01	0,25	binary	28,29	76,19	41,26	20,40	94,24	33,54	55,73	96,99	70,78	40,49	84,01	54,64	50,05
253	30	50	-0,8	0,01	0,25	tf	28,96	96,87	44,59	27,16	94,24	42,16	92,61	88,86	90,69	53,66	84,99	65,79	60,81
254	30	50	-0,8	0,01	0,9	tf*idf	28,61	76,46	41,64	24,59	94,63	39,04	77,43	88,17	82,45	53,83	84,26	65,69	57,20
255	30	50	-0,8	0,00001	0,25	tf*idf	29,00	96,81	44,63	25,02	94,09	39,52	92,95	88,90	90,88	49,82	84,09	62,57	59,40
256	30	50	0,2	0,01	0,25	binary	28,37	20,62	23,88	96,35	34,77	51,10	92,38	79,47	85,44	61,70	57,66	59,61	55,01
257	30	50	0,8	0,01	0,01	tf	28,78	20,53	23,96	96,20	6,92	12,92	92,43	63,59	75,34	88,70	18,27	30,30	35,63
258	30	50	0,8	0,00001	0,9	tf*idf	28,74	36,84	32,29	26,89	54,03	35,91	92,23	46,14	61,51	51,38	75,29	61,08	47,70
259	30	50	0,8	0,00001	0,01	tf	28,35	20,75	23,97	96,43	5,27	9,99	92,02	63,78	75,34	88,11	18,38	30,41	34,93
260	30	50	0,8	0,00001	0,01	tf*idf	28,15	20,09	23,45	96,23	5,77	10,88	92,76	63,55	75,43	88,86	11,16	19,83	32,40
261	50	50	-0,8	0,01	0,25	binary	28,24	76,56	41,26	19,06	94,42	31,72	67,87	96,03	79,53	43,50	84,66	57,47	52,50
262	50	50	-0,8	0,01	0,25	tf	28,83	96,89	44,44	27,84	94,09	42,97	92,29	88,92	90,57	53,45	84,29	65,41	60,85
263	50	50	0,2	0,00001	0,25	binary	28,07	20,65	23,80	96,62	34,46	50,80	92,46	72,00	80,96	65,86	75,42	70,32	56,47
264	500	50	-0,8	0,01	0,25	binary	28,84	76,56	41,90	32,41	94,92	48,32	67,14	96,01	79,02	40,58	84,15	54,75	56,00
265	500	50	-0,8	0,01	0,9	binary	28,90	76,10	41,89	27,33	74,52	40,00	55,03	96,89	70,20	43,36	84,00	57,20	52,32
266	30	50	-0,8	0,01	0,25	tf*idf	26,43	96,93	41,53	24,03	94,64	38,32	78,79	96,95	86,93	53,01	84,94	65,28	58,02
267	50	50	-0,8	0,01	0,25	tf*idf	26,89	96,94	42,11	27,51	94,74	42,63	78,17	96,85	86,51	53,63	84,62	65,65	59,23
268	30	50	-0,8	0,00001	0,01	binary	24,64	96,48	39,26	17,16	94,53	29,04	67,14	96,57	79,21	36,08	84,54	50,57	49,52
269	30	50	-0,8	0,00001	0,01	tf	24,08	96,99	38,58	22,98	94,78	37,00	84,64	96,23	90,06	40,87	84,74	55,14	55,19
270	30	50	0,8	0,00001	0,25	tf*idf	24,32	36,45	29,17	63,53	34,04	44,33	92,57	71,23	80,51	68,41	20,47	31,51	46,38
271	50	1	-0,8	0,00001	0,01	binary	24,85	96,93	39,55	25,68	94,81	40,41	44,74	96,30	61,10	43,81	84,91	57,80	49,72
272	50	50	-0,8	0,00001	0,01	binary	24,94	96,00	39,60	17,24	94,63	29,17	67,26	96,71	79,34	36,47	84,08	50,88	49,75
273	30	1	-0,8	0,01	0,01	tf*idf	23,82	96,57	38,22	52,04	94,70	67,17	55,42	96,60	70,43	49,28	84,57	62,27	59,52
274	30	1	-0,8	0,00001	0,01	tf*idf	23,70	96,96	38,09	59,88	94,77	73,39	55,84	96,07	70,63	49,35	84,77	62,38	61,12
275	500	50	-0,8	0,01	0,01	tf*idf	23,99	96,79	38,45	21,74	94,39	35,34	92,21	96,49	94,30	40,92	84,99	55,24	55,83
276	50	1	-0,8	0,9	0,9	binary	23,42	96,40	37,68	41,56	94,89	57,80	38,47	96,88	55,08	25,27	84,42	38,89	47,36
277	500	1	-0,8	0,01	0,01	binary	23,57	96,70	37,90	21,25	94,48	34,70	42,14	96,45	58,65	43,10	84,56	57,10	47,09
278	50	50	-0,8	0,01	0,9	binary	22,64	76,88	34,98	36,64	74,24	49,07	67,63	96,52	79,53	43,54	84,68	57,51	55,27
279	30	1	-0,8	0,01	0,01	tf	21,48	96,84	35,16	46,13	94,47	61,99	59,56	96,19	73,56	49,73	85,00	62,75	58,37
280	30	1	-0,8	0,9	0,25	tf	21,83	96,25	35,58	22,65	94,38	36,53	42,87	96,96	59,45	38,87	84,18	53,18	46,19
281	30	1	-0,8	0,9	0,9	tf	21,10	96,28	34,61	32,42	74,15	45,12	53,12	88,38	66,36	53,83	84,15	65,66	52,94
282	30	1	-0,8	0,00001	0,01	binary	21,76	96,12	35,49	38,08	94,89	54,35	42,67	96,98	59,27	46,87	84,77	60,36	52,37
283	30	1	-0,8	0,00001	0,01	tf	21,84	96,49	35,62	46,90	94,74	62,74	59,54	96,75	73,72	49,04	84,50	62,06	58,53
284	30	1	-0,8	0,9	0,25	tf*idf	20,76	96,82	34,19	29,27	94,66	44,71	35,66	96,85	52,13	34,97	84,19	49,42	45,11
285	30	1	-0,8	0,9	0,9	tf*idf	20,56	76,42	32,40	38,20	94,48	54,40	38,61	96,98	55,23	49,03	84,55	62,07	51,03
286	30	50	-0,8	0,01	0,01	tf	20,35	96,77	33,63	19,79	94,87	32,75	84,66	96,89	90,37	40,66	84,04	54,81	52,89
287	30	50	0,2	0,01	0,01	tf	20,78	20,96	20,87	96,65	3,74	7,21	92,75	88,29	90,46	71,29	75,66	73,41	47,99
288	30	50	0,2	0,00001	0,25	binary	20,67	20,87	20,77	96,10	6,94	12,95	92,81	88,46	90,58	61,13	57,17	59,09	45,85
289	30	50	0,2	0,00001	0,01	tf	20,11	20,34	20,23	96,03	5,96	11,23	92,42	88,66	90,50	71,44	75,92	73,61	48,89
290	30	50	0,8	0,00001	0,25	tf	20,04	20,61	20,32	96,63	20,21	33,43	92,40	71,89	80,86	76,70	47,32	58,53	48,29
291	50	50	-0,8	0,9	0,9	binary	20,55	76,95	32,43	19,74	54,65	29,01	28,31	96,16	43,75	51,91	75,05	61,37	41,64
292	30	1	-0,8	0,01	0,01	binary	19,60	96,96	32,62	34,87	94,46	50,94	42,05	96,29	58,54	46,11	84,57	59,68	50,44
293	500	1	-0,8	0,9	0,25	binary	19,63	96,91	32,65	20,14	94,23	33,18	33,96	96,20	50,20	30,88	84,75	45,27	40,33
294	30	1	-0,8	0,9	0,9	binary	18,42	76,48	29,69	20,09	94,99	33,17	31,56	96,42	47,55	51,41	75,79	61,26	42,92
295	30	1	-0,8	0,9	0,01	tf	18,21	96,08	30,62	20,44	94,90	33,63	34,72	96,84	51,12	37,63	84,04	51,98	41,84
296	30	50	-0,8	0,9	0,9	binary	18,96	56,36	28,38	20,50	74,78	32,18	28,89	96,66	44,48	51,79	75,14	61,32	41,59
297	50	50	-0,8	0,01	0,01	binary	18,40	96,89	30,93	17,07	74,51	27,78	67,19	96,93	79,36	36,21	84,31	50,66	47,18
298	30	50	-0,8	0,01	0,01	binary	17,48	96,32	29,59	17,64	94,78	29,74	67,05	96,66	79,18	36,93	84,61	51,42	47,48
299	500	1	-0,8	0,9	0,9	binary	17,82	96,24	30,07	24,46	94,93	38,90	40,16	96,12	56,65	40,72	84,81	55,02	45,16
300	30	50	-0,8	0,9	0,25	tf	16,00	96,57	27,45	21,71	94,99	35,35	31,79	88,56	46,78	32,21	84,65	46,66	39,06
301	30	50	-0,8	0,9	0,9	tf	16,31	96,81	27,92	19,08	94,47	31,75	34,85	88,80	50,05	41,48	75,99	53,66	40,85
302	500	1	-0,8	0,9	0,01	binary	16,07	96,80	27,56	20,44	94,50	33,61	33,52	96,08	49,70	26,75	84,20	40,60	37,87
303	30	1	-0,8	0,9	0,01	tf*idf	20,96	96,43	34,43	22,42	94,22	36,22	33,93	96,05	50,15	36,94	84,33	51,37	43,05
304	30	50	-0,8	0,9	0,25	tf*idf	20,47	96,91	33,80	20,10	94,16	33,12	31,00	88,72	45,95	31,38	75,57	44,35	39,31
305	30	50	0,2	0,01	0,01	tf*idf	21,00	20,96	20,98	96,88	4,32	8,27	92,44	88,43	90,39	59,59	75,98	66,79	46,61
306	30	50	0,2	0,00001	0,01	tf*idf	20,95	20,64	20,80	96,33	0,65	1,30	92,81	88,24	90,47	59,42	75,49	66,50	44,76
307	30	50	-0,8	0,9	0,9	tf*idf	19,51	76,75	31,11	21,71	94,42	35,31	33,23	96,78	49,47	32,22	84,06	46,58	40,62
308	30	50	-0,8	0,9	0,01	tf	19,73	96,98	32,79	17,06	94,57	28,90	33,00	96,66	49,20	32,55	84,38	46,98	39,47
309	30	50	-0,8	0,9	0,01	tf*idf	18,32	96,62	30,81	20,44	94,08	33,59	31,06	96,08	46,95	32,16	84,60	46,61	39,49
310	30	50	0,2	0,00001	0,9	tf	17,77	20,18	18,90	29,87	34,74	32,13	92,40	79,01	85,18	59,48	75,09	66,38	50,65
311	50	1	-0,8	0,9	0,25	binary	17,89	96,29	30,18	17,94	94,65	30,16	28,70	96,38	44,23	23,03	84,02	36,15	35,18
312	50	1	-0,8	0,9	0,01	binary	17,78	96,52	30,03	17,08	94,44	28,93	28,40	96,33	43,86	30,06	84,44	44,34	36,79
313	30	1	-0,8	0,9	0,01	binary	16,01	96,08	27,45	17,27	94,38	29,19	28,54	96,22	44,02	27,96	84,12	41,97	35,66
314	30	1	-0,8	0,9	0,25	binary	15,41	96,99	26,60	17,79	94,60	29,95	28,26	96,54	43,72	23,66	84,46	36,96	34,31
315	30	50	-0,8	0,9	0,01	binary	15,85	96,49	27,23	15,55	94,75	26,72	28,41	96,05	43,84	23,72	84,65	37,05	33,71
316	50	50	-0,8	0,9	0,01	binary	15,24	97,00	26,34	15,40	94,48	26,49	28,17	96,87	43,65	23,62	84,01	36,87	33,34
317	30	50	-0,8	0,9	0,25	binary	14,45	76,68	24,31	19,15	94,92	31,87	28,81	96,64	44,38	22,28	84,27	35,25	33,95
318	50	50	-0,8	0,9	0,25	binary	13,94	76,80	23,59	19,12	94,15	31,79	28,21	96,68	43,68	21,16	84,94	33,88	33,24
319	30	1	0,8	0,9	0,9	tf*idf	4,95	5,54	5,23	96,98	20,57	33,94	92,99	63,76	75,65	88,09	20,82	33,68	37,13
320	30	50	0,8	0,01	0,25	tf	0,69	4,55	1,20	96,23	4,27	8,18	92,12	71,19	80,32	68,68	20,39	31,45	30,29
321	30	50	0,8	0,01	0,9	binary	0,43	0,36	0,39	96,53	0,65	1,29	92,71	46,28	61,74	88,79	11,29	20,03	20,86
322	30	50	0,8	0,01	0,01	tf*idf	1,47	1,82	1,62	96,10	3,00	5,82	92,94	63,99	75,80	88,69	11,64	20,58	25,95
323	50	50	0,8	0,01	0,9	binary	4,07	4,78	4,40	96,61	4,75	9,05	92,52	54,24	68,38	88,90	39,94	55,12	34,24

F) Programa Batch com os principais comandos para executar o classificador por Indução de Regras

PRIMEIRA EXECUÇÃO – C/ARQUIVO DE TREINAMENTO

```
java mkdict 500 dirik.dit

java vectorize alimen ve1ctr.vec
java vectorize equipa ve2ctr.vec
java vectorize hext   ve3ctr.vec
java vectorize honora ve4ctr.vec
```

SEGUNDA EXECUÇÃO – C/ARQUIVO DE TESTE

```
java vectorize alimen ve1cte.vec
riktext -t ve1cte.vec dirik.dit alimen ve1ctr.vec >cl1ctet.txt

java vectorize equipa ve2cte.vec
riktext -t ve2cte.vec dirik.dit equipa ve2ctr.vec >cl2ctet.txt

java vectorize hext ve3cte.vec >x.txt
riktext -t ve3cte.vec dirik.dit hext   ve3ctr.vec >cl3ctet.txt

java vectorize honora ve4cte.vec
riktext -t ve4cte.vec dirik.dit honora ve4ctr.vec >cl4ctet.txt
```


G) Resultados encontrados com o classificador por Indução de Regras

						Alimentação			Equiparação			Horas Extras			Honorários			
Num reg	Num Pal	Lim freq	Tip test	Trade off	Sele ção	prec	cob	med F	prec	cob	med F	prec	cob	med F	prec	cob	med F	med F media
1	30	0	1	0	0	85,26	1,26	2,49	82,50	25,30	38,72	85,57	77,43	81,30	75,59	82,40	78,85	50,34
2	30	0	1	0	1	85,76	1,89	3,70	82,77	25,48	38,97	85,36	77,78	81,39	75,99	82,93	79,31	50,84
3	30	0	1	0	6	85,33	2,49	4,83	82,50	25,75	39,25	85,39	77,83	81,44	75,97	82,16	78,94	51,12
4	30	0	2	0	0	85,51	61,55	71,58	57,28	45,45	50,69	85,62	77,95	81,61	75,81	82,32	78,93	70,70
5	30	0	2	0	1	85,41	1,85	3,63	57,52	45,47	50,79	85,08	77,82	81,29	75,55	82,63	78,93	53,66
6	30	0	2	0	6	85,28	61,71	71,60	57,96	45,30	50,85	85,30	77,69	81,32	75,54	82,01	78,65	70,61
7	30	1	1	0	0	85,55	6,59	12,23	82,81	45,07	58,37	75,97	60,75	67,51	75,69	82,98	79,17	54,32
8	30	1	1	0	1	85,52	6,31	11,76	82,80	45,35	58,60	75,03	60,15	66,77	75,38	82,78	78,91	54,01
9	30	1	1	0	6	85,95	1,59	3,13	82,86	45,86	59,04	75,15	60,23	66,87	75,07	82,01	78,39	51,86
10	30	2	1	0	0	85,55	2,98	5,77	82,65	25,72	39,23	85,92	77,27	81,37	75,88	82,62	79,11	51,37
11	30	2	1	0	1	85,87	0,63	1,25	82,11	25,01	38,34	85,91	77,05	81,24	75,90	82,41	79,02	49,96
12	30	2	1	0	6	85,29	3,89	7,45	82,65	25,74	39,26	85,92	77,05	81,24	75,40	82,07	78,59	51,64
13	30	2	2	0	0	85,71	61,19	71,40	57,65	45,35	50,77	85,51	77,95	81,56	75,77	82,78	79,12	70,71
14	30	2	2	0	1	85,49	61,91	71,81	57,17	45,08	50,41	85,40	77,35	81,18	75,55	82,76	78,99	70,60
15	30	2	2	0	6	86,00	61,43	71,67	57,06	45,99	50,93	85,96	77,57	81,55	75,93	82,30	78,99	70,78
16	50	0	1	0	0	85,43	41,40	55,77	82,88	85,53	84,19	85,28	77,15	81,01	75,12	82,30	78,55	74,88
17	50	0	1	0	1	85,65	41,32	55,75	82,18	85,65	83,88	85,82	77,15	81,26	75,26	82,11	78,54	74,86
18	50	0	1	0	6	85,28	41,59	55,91	82,19	85,50	83,81	85,52	77,34	81,23	75,62	82,21	78,78	74,93
19	50	0	1	4	0	85,62	41,70	56,09	82,13	85,65	83,85	85,08	77,37	81,04	75,37	82,13	78,60	74,89
20	50	0	1	4	1	85,89	41,57	56,03	82,06	85,58	83,78	85,90	77,92	81,72	75,01	82,50	78,58	75,03
21	50	0	1	4	6	85,73	41,01	55,48	82,58	85,61	84,07	85,07	77,45	81,08	75,70	82,59	78,99	74,91
22	50	0	1	9	0	85,78	41,41	55,86	82,18	85,54	83,83	85,24	77,92	81,42	75,33	82,66	78,82	74,98
23	50	0	1	9	1	85,71	41,99	56,37	82,58	85,72	84,12	85,20	77,07	80,94	75,33	82,90	78,93	75,09
24	50	0	1	9	6	85,96	41,52	55,99	82,40	85,73	84,03	85,71	77,83	81,58	75,82	82,20	78,88	75,12
25	50	0	2	0	0	85,72	41,54	55,96	82,23	85,19	83,68	85,12	77,76	81,27	75,54	82,48	78,86	74,94
26	50	0	2	4	0	85,36	61,55	71,52	82,29	85,61	83,92	85,05	77,20	80,93	75,27	82,00	78,49	78,72
27	50	0	2	9	0	85,92	61,23	71,51	82,48	85,38	83,90	85,36	77,51	81,25	75,43	82,16	78,65	78,83
28	50	1	1	0	0	85,29	41,84	56,14	82,00	85,91	83,91	75,67	60,66	67,34	75,12	82,66	78,71	71,52
29	50	1	1	4	0	85,13	41,68	55,96	82,66	85,22	83,92	77,55	85,01	81,11	75,50	82,96	79,06	75,01
30	50	1	1	9	0	85,65	41,68	56,07	82,05	85,38	83,68	77,06	85,38	81,01	75,03	82,40	78,54	74,83
31	50	1	2	0	0	85,88	41,89	56,31	82,99	85,16	84,06	85,81	85,44	85,62	75,80	82,66	79,08	76,27
32	50	1	2	4	0	85,85	41,54	55,99	82,20	85,92	84,02	77,14	85,70	81,19	75,21	82,70	78,78	75,00
33	50	1	2	9	0	85,93	41,65	56,10	82,26	85,26	83,74	85,02	77,52	81,10	75,70	82,66	79,03	74,99
34	50	2	1	0	0	85,92	41,95	56,37	82,16	85,18	83,64	85,58	77,24	81,20	75,30	82,20	78,60	74,95
35	50	2	1	4	0	85,50	41,78	56,13	82,96	85,28	84,10	85,49	77,53	81,32	75,02	82,11	78,40	74,99
36	50	2	1	9	0	85,76	41,22	55,68	82,63	85,90	84,23	85,92	77,64	81,57	75,69	82,70	79,04	75,13
37	50	2	2	0	0	85,69	41,83	56,21	82,46	85,80	84,09	85,81	77,28	81,33	75,48	82,29	78,74	75,09
38	50	2	2	4	0	85,05	41,30	55,60	82,10	85,85	83,94	85,86	77,15	81,27	75,39	82,61	78,84	74,91
39	50	2	2	9	1	85,98	61,13	71,46	82,65	85,22	83,91	85,41	77,59	81,31	75,46	82,42	78,78	78,87
40	50	2	2	9	6	85,57	61,32	71,44	82,61	85,54	84,05	85,25	77,28	81,07	75,65	82,41	78,89	78,86
41	100	0	1	0	0	85,75	61,84	71,86	82,41	85,63	83,99	85,36	77,12	81,04	75,44	82,25	78,70	78,90
42	100	0	1	0	1	85,97	61,51	71,71	82,54	85,13	83,82	85,39	77,28	81,13	75,63	82,49	78,91	78,89
43	100	0	1	0	6	85,83	61,04	71,34	82,21	85,02	83,59	85,74	77,85	81,60	75,10	82,50	78,63	78,79
44	100	0	1	4	0	85,56	61,41	71,50	82,35	85,67	83,98	85,08	77,72	81,23	75,79	82,62	79,06	78,94
45	100	0	1	4	1	85,55	61,75	71,73	82,70	85,31	83,98	85,31	78,00	81,49	75,14	82,05	78,45	78,91
46	100	0	1	4	6	85,65	61,87	71,84	82,79	85,29	84,02	85,25	77,81	81,36	75,75	82,85	79,14	79,09
47	100	0	1	9	0	85,68	61,13	71,35	82,99	85,66	84,30	85,09	77,91	81,34	75,56	82,88	79,05	79,01
48	100	0	1	9	1	85,27	61,78	71,65	82,18	85,39	83,76	85,17	77,54	81,17	75,49	82,62	78,89	78,87
49	100	0	1	9	6	85,86	61,25	71,50	82,36	85,71	84,00	85,13	77,91	81,36	75,70	82,59	79,00	78,96
50	100	0	2	0	0	85,14	61,76	71,59	82,68	85,07	83,86	85,45	77,19	81,11	75,94	82,05	78,88	78,86
51	100	0	2	0	1	85,79	61,31	71,51	82,22	85,35	83,75	85,33	77,14	81,03	75,89	82,93	79,26	78,89
52	100	0	2	4	0	85,04	61,00	71,04	82,25	85,50	83,84	85,85	77,96	81,72	75,65	82,30	78,84	78,86
53	100	0	2	9	0	85,41	61,73	71,66	82,76	85,74	84,23	85,75	77,25	81,28	75,12	82,86	78,80	78,99
54	100	1	1	0	0	85,35	41,49	55,84	82,36	85,95	84,12	75,80	60,06	67,01	75,13	82,55	78,66	71,41
55	100	1	1	4	0	85,26	41,59	55,91	82,97	85,03	83,98	77,32	85,79	81,33	75,03	82,40	78,54	74,94

56	100	1	1	9	0	85,36	61,72	71,64	82,55	85,09	83,80	77,06	85,66	81,13	75,75	82,62	79,03	78,90
57	100	1	2	0	0	85,38	61,53	71,52	82,17	85,26	83,69	85,14	85,07	85,10	75,64	82,13	78,75	79,77
58	100	1	2	4	0	85,39	41,23	55,61	82,80	86,00	84,37	77,33	85,42	81,17	75,84	82,85	79,19	75,08
59	100	1	2	9	0	85,16	41,18	55,52	82,39	85,24	83,79	77,10	85,66	81,16	75,39	82,34	78,71	74,79
60	100	2	1	0	0	85,77	61,82	71,85	82,33	85,47	83,87	85,56	77,87	81,53	75,73	82,03	78,75	79,00
61	100	2	1	4	0	85,71	61,37	71,53	82,72	85,89	84,28	85,25	77,90	81,41	75,17	82,47	78,65	78,97
62	100	2	1	9	0	85,64	61,21	71,40	82,13	85,03	83,56	85,86	77,74	81,60	75,25	82,75	78,82	78,84
63	100	2	2	0	0	85,53	41,03	55,46	82,81	85,90	84,33	85,66	77,40	81,32	75,53	82,88	79,03	75,04
64	100	2	2	4	0	85,06	41,18	55,49	82,41	85,94	84,14	85,84	77,05	81,21	75,89	82,82	79,20	75,01
65	100	2	2	9	0	85,30	41,58	55,91	82,45	85,06	83,73	85,74	77,53	81,43	75,44	82,91	79,00	75,02
66	30	0	2	9	1	65,66	61,81	63,68	32,80	65,16	43,64	85,41	77,77	81,41	75,26	82,06	78,52	66,81
67	30	2	2	4	0	65,98	61,59	63,71	57,34	45,73	50,88	85,12	77,73	81,25	75,72	82,14	78,80	68,66
68	30	2	2	4	6	65,69	61,72	63,64	57,83	45,05	50,65	85,55	77,03	81,07	75,96	82,23	78,97	68,58
69	30	2	2	9	0	65,60	61,24	63,35	32,09	65,29	43,03	85,67	77,38	81,32	75,20	82,43	78,65	66,58
70	30	2	2	9	6	65,18	61,04	63,04	32,77	65,79	43,75	85,21	77,32	81,07	75,50	82,99	79,07	66,73
71	30	1	2	0	0	60,58	41,35	49,15	82,09	5,67	10,60	75,51	60,10	66,93	75,52	82,73	78,96	51,41
72	30	1	2	0	1	60,09	41,47	49,07	82,53	4,60	8,72	75,24	60,71	67,20	75,45	82,25	78,70	50,92
73	30	1	2	0	6	60,45	41,80	49,42	82,05	5,25	9,88	75,31	60,02	66,80	75,78	82,20	78,86	51,24
74	30	1	2	4	0	60,89	41,11	49,09	82,21	0,39	0,77	77,66	86,91	81,58	75,52	82,12	78,68	52,53
75	30	1	2	4	6	60,34	41,26	49,01	82,99	5,67	10,61	77,34	85,51	81,22	75,32	82,07	78,55	54,85
76	30	2	2	4	1	42,75	61,57	50,46	57,40	45,41	50,71	85,64	77,38	81,30	75,40	82,61	78,84	65,33
77	30	2	2	9	1	42,82	61,81	50,59	32,80	65,18	43,64	85,27	77,27	81,07	75,33	82,08	78,56	63,47
78	30	0	1	4	0	35,83	61,26	45,21	82,89	2,85	5,51	85,91	77,21	81,33	75,53	82,71	78,96	52,75
79	30	0	1	4	1	35,33	61,36	44,84	82,82	0,56	1,12	85,45	77,81	81,45	75,33	82,03	78,54	51,49
80	30	0	1	4	6	35,78	61,33	45,20	82,45	6,05	11,28	85,02	77,42	81,04	75,16	82,89	78,84	54,09
81	30	0	2	4	0	35,45	61,58	45,00	57,25	45,56	50,74	85,42	77,43	81,23	75,30	82,49	78,73	63,93
82	30	0	2	4	1	35,63	61,93	45,23	57,04	45,44	50,58	85,66	77,22	81,22	75,37	82,63	78,83	63,97
83	30	0	2	4	6	35,44	61,12	44,87	57,56	45,58	50,88	85,24	77,30	81,07	75,87	82,62	79,10	63,98
84	30	0	2	9	0	35,59	61,61	45,12	32,48	65,83	43,50	85,70	77,65	81,48	75,83	82,09	78,84	62,23
85	30	0	2	9	6	35,27	61,68	44,88	32,54	65,08	43,38	85,11	77,29	81,01	75,31	82,81	78,88	62,04
86	30	1	1	4	0	35,05	61,92	44,77	82,41	2,16	4,20	77,06	85,95	81,26	75,77	82,15	78,83	52,27
87	30	1	1	4	1	35,37	61,16	44,82	82,26	1,79	3,51	77,10	85,07	80,89	75,01	82,89	78,75	51,99
88	30	1	1	4	6	35,69	61,69	45,22	82,72	1,02	2,02	77,85	85,24	81,38	75,75	82,78	79,11	51,93
89	30	2	1	4	0	35,88	61,48	45,31	82,50	0,86	1,69	85,08	77,26	80,98	75,05	82,46	78,58	51,64
90	30	2	1	4	1	35,34	61,92	45,00	82,98	0,65	1,29	85,82	77,51	81,45	75,16	82,97	78,87	51,65
91	30	2	1	4	6	35,84	61,22	45,21	82,16	2,02	3,94	85,51	77,15	81,11	75,39	82,39	78,73	52,25
92	30	1	2	9	0	29,16	61,83	39,63	39,30	65,94	49,24	77,79	85,61	81,51	75,79	82,29	78,90	62,32
93	30	1	2	9	6	29,92	61,17	40,19	39,62	65,47	49,37	77,01	85,59	81,08	75,56	82,32	78,79	62,36
94	30	1	2	9	1	23,85	81,98	36,96	39,99	65,62	49,69	77,40	85,04	81,04	75,78	82,68	79,08	61,69
95	30	1	2	4	1	21,13	61,68	31,47	82,29	4,20	8,00	77,26	85,70	81,26	75,72	82,29	78,87	49,90
96	30	0	1	9	0	18,49	61,18	28,39	32,59	65,29	43,48	85,69	77,90	81,61	75,35	82,36	78,70	58,05
97	30	0	1	9	1	18,92	61,03	28,88	32,45	65,10	43,31	85,34	77,84	81,41	75,78	82,94	79,20	58,20
98	30	0	1	9	6	18,71	61,91	28,73	32,16	65,15	43,06	85,67	77,90	81,60	75,27	82,54	78,74	58,03
99	30	2	1	9	0	8,52	81,95	15,43	32,17	65,03	43,05	85,73	77,29	81,29	75,61	82,23	78,78	54,64
100	30	2	1	9	1	8,45	81,42	15,31	32,72	65,56	43,65	85,09	77,47	81,10	75,83	82,53	79,04	54,78
101	30	2	1	9	6	8,48	81,88	15,37	32,89	65,22	43,73	85,28	77,98	81,47	75,12	82,97	78,85	54,85
102	30	1	1	9	0	6,96	61,22	12,49	42,72	45,88	44,24	77,99	85,78	81,70	75,54	82,20	78,73	54,29
103	30	1	1	9	1	6,27	61,32	11,38	42,68	45,49	44,04	77,74	85,96	81,64	75,41	82,17	78,65	53,93
104	30	1	1	9	6	6,38	61,78	11,57	42,76	45,99	44,32	77,65	85,88	81,56	75,88	82,74	79,16	54,15

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo