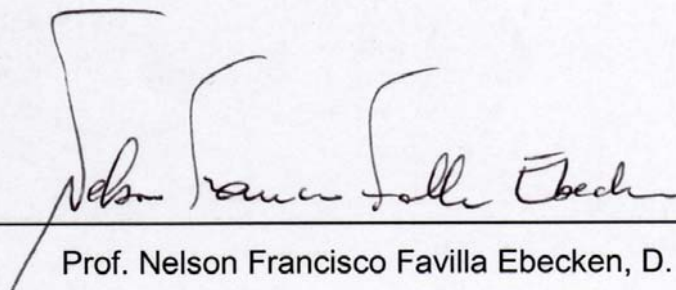


IMPLEMENTAÇÃO DE METODOLOGIA DE CATEGORIZAÇÃO DE TEXTOS
CIENTÍFICOS

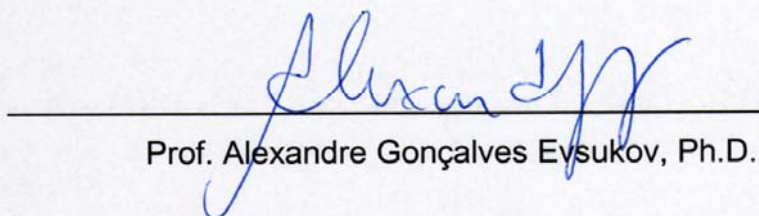
Marco Aurélio Ribeiro Dantas

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.

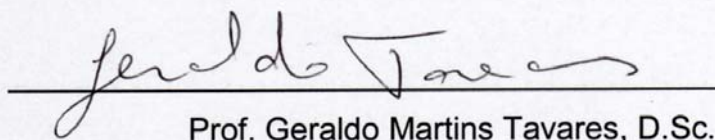
Aprovada por:



Prof. Nelson Francisco Favilla Ebecken, D. Sc.



Prof. Alexandre Gonçalves Eysukov, Ph.D.



Prof. Geraldo Martins Tavares, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2007

DANTAS, MARCO AURÉLIO RIBEIRO

Implementação de metodologia de categorização de textos científicos [Rio de Janeiro] 2007.

X, 86 p. 29,7 cm (COPPE/UFRJ,M.Sc., Engenharia Civil, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Categorização

2. Text Mining

I. COPPE/UFRJ II. Título (série)

DEDICATÓRIA

A Deus por me conceder persistência e amor à vida,
A todos que, direta ou indiretamente, me ajudaram nesta empreitada.

AGRADECIMENTOS

Agradeço em primeiro a Deus, sem ele nada é possível.

Agradeço ao meu orientador, Prof. Nelson F. F. Ebecken, pelo auxílio, dedicação, paciência e confiança na composição deste trabalho.

Ao meu chefe, amigo e grande incentivador, Prof. Geraldo Tavares.

A minha noiva, Andréia Dutra Fraguas, pelo carinho nos momentos difíceis.

As minhas colegas Cláudia Paes, Marina Pires e ao meu colega Antonio Anddre Serpa da Silva, com todo o apoio técnico e incentivo nos momentos difíceis.

Aos meus colegas de trabalho no LEV. pelo incentivo e motivação dadas por meus amigos e familiares.

Aos amigos e familiares, pelo carinho e, que de alguma forma torcem pelo meu sucesso.

A Secretaria do Programa e a Secretaria do Núcleo de Transferência de Tecnologia, pelo suporte administrativo prestado.

Ao CNPq pelo suporte financeiro.

Enfim, essa dissertação não é só minha, é de todos vocês! Sem vocês esta conquista não seria possível.

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

IMPLEMENTAÇÃO DE METODOLOGIA DE CATEGORIZAÇÃO DE TEXTOS CIENTÍFICOS

Marco Aurélio Ribeiro Dantas

Dezembro/2007

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Esta dissertação apresenta uma metodologia de categorização de textos para ser utilizada em acervos digitais de laboratórios científicos. Foram considerados os idiomas português e inglês. A manipulação e preparação de dados utiliza o sistema Aîuri e os categorizadores linear e bayesiano. Efetuou-se a sua utilização em um caso de laboratório de engenharia, comentando-se as características e o desempenho dos resultados obtidos. A estratégia está pronta para ser utilizada diretamente em aplicações científicas, podendo ser facilmente expandida para incluir novas facilidades.

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

IMPLEMENTATION OF METHODOLOGY OF CATEGORIZATION OF SCIENTIFIC TEXTS

Marco Aurélio Ribeiro Dantas

December/2007

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This dissertation presents a methodology of categorization of texts to be used in digital collections of scientific laboratories. The Portuguese and English languages were considered. The manipulation and preparation of data uses the system Aiuri and the linear and bayesian categorizators. The uses occurred in a case of engineering laboratory, being commented on the characteristics and the acting of the obtained results. The strategy is ready to be used directly in scientific applications, could be easily expanded to include new facilities.

SUMÁRIO

LISTA DE FIGURAS.....	IX
LISTA DE TABELAS.....	X
LISTA DE ABREVIACOES E SIGLAS.....	XI
1 INTRODUO.....	1
1.1 FORMULAO DA SITUAO-PROBLEMA.....	3
1.2 OBJETIVOS DA TESE	4
1.2.1 Objetivo Geral.....	4
1.2.2 Objetivos Especficos.....	4
1.3 RELEVNCIA	5
2 METODOLOGIA	7
3 REVISO BIBLIOGRFICA	9
3.1 INFORMAO	9
3.2 RECUPERAO DE INFORMAOES	11
3.3 TCNICAS DE MINERAO DE TEXTOS	12
3.3.1 Introduo e Consideraes Gerais	12
3.3.2 Abordagens dos dados utilizados em Minerao de Texto	12
3.3.2.1 Anlise semntica dos dados	13
3.3.2.2 Anlise estatstica baseada em freqncia	13
3.3.3 Etapas do processo de Minerao de Textos.....	13
3.3.4 Principais tcnicas de minerao de textos.....	14
3.4 CATEGORIZAO DE TEXTO.....	15
3.4.1 Categorizao Manual.....	16
3.4.2 Categorizao Automtica.....	16
3.5 MTODOS DE CATEGORIZAO AUTOMTICA USADOS EM APRENDIZADO DE MQUINA.....	19
3.6 METODOLOGIA DO PROCESSO DE CATEGORIZAO	25
3.6.1 Coleta dos textos	25
3.6.2 Pr-processamento.....	25
3.6.2.1 Case Folding.....	26
3.6.2.2 Identificao de Termos.....	26

4	ESCOLHA DA METODOLOGIA DE CATEGORIZAÇÃO DE TEXTOS, SOFTWARES E ALGORITMOS A SEREM UTILIZADOS	31
4.1	ESCOLHA DA METODOLOGIA, ALGORITMOS E SOFTWARES.....	31
4.2	ALGORITMOS UTILIZADOS.....	31
4.3	Descrição da metodologia de aplicação do conjunto	31
5	APLICAÇÃO AO CASO DO LABORATÓRIO DE ENERGIA DOS VENTOS.....	32
5.1	COLETA DOS TEXTOS	32
5.2	PRÉ-PROCESSAMENTO	36
5.2.1	CASE FOLDING	36
5.2.2	IDENTIFICAÇÃO DE TERMOS.....	37
5.2.5	SELEÇÃO DAS CARACTERÍSTICAS.....	37
5.2.5.1	Seleção por frequência relativa	38
5.2.5.2	Seleção por frequência do documento	38
5.2.5.3	Produto da frequência do termo pela frequência inversa de documentos	39
5.2.6	VETORES ESPARSOS.....	39
5.2.6.1	Vetores rotulados e não-rotulados.....	40
5.2.6.2	Índice Invertido	40
6	PROCESSAMENTO DOS DADOS	41
6.1	CLASSIFICADOR BAYESIANO	44
6.2	CLASSIFICADOR LINEAR.....	44
7	AVALIAÇÃO DOS RESULTADOS	45
7.1	PRECISÃO	45
7.2	REVOCAÇÃO.....	45
7.3	MÉDIA F	45
7.4	MATRIZ DE CONFUSÃO	46
7.5	RESULTADOS PARA A BASE EM IDIOMA PORTUGUÊS.....	46
7.5.1	Classe Data Mining.....	46
7.5.2	Classe Eficiência Energética	47
7.5.3	Classe Energia	48
7.5.4	Classe SIIF	49
7.5.5	Classe Equipamento Elétrico.....	50
7.5.6	Classe Gerenciamento Energético	50
7.5.7	Classe Pesquisa e Desenvolvimento.....	51
7.5.8	Classe Energia Eólica.....	52
7.6	RESULTADOS PARA A BASE EM IDIOMA INGLÊS	53
7.6.1	Classe Data Mining.....	53

7.6.2 Classe Eficiência Energética	54
7.6.3 Classe Energia	55
7.6.4 Classe Educação	55
7.6.5 Classe Energia Distribuída	56
7.6.6 Classe Manutenção	57
7.6.7 Classe Pesquisa e Desenvolvimento.....	58
7.6.8 Classe Energia Eólica.....	58
7.6.9 Classe Geração Distribuída	59
7.6.10 Classe Mercado	60
7.6.11 Classe Mudanças Climáticas.....	61
7.6.12 Classe Qualidade de Energia	61
7.6.13 Classe Renováveis	62
7.6.14 Classe Sistemas de Potência	63
7.6.15 Classe Termelétrica.....	64
7.6.16 Classe Tecnologia da Informação	65
7.6.17 Classe Transmissão	65
8 CONCLUSÕES.....	67
8.1 SUGESTÕES PARA TRABALHOS POSTERIORES	68
REFERÊNCIAS BIBLIOGRÁFICAS.....	69
APÊNDICE 1.....	72
APÊNDICE 2.....	77

LISTA DE FIGURAS

Figura 1 - Principais etapas do processo de Text Mining.....	15
Figura 2 - Processo de categorização.....	18
Figura 3 - Esquema da etapa de aprendizado.....	19
Figura 4 - Esquema da Etapa de classificação.....	19
Figura 5 - Exemplo do uso de um conjunto de pesos para determinar a contagem para o documento.....	25
Figura 6 - Esquema da etapa de pré-processamento.....	26
Figura 7 - Ilustração de Bag Of Words.....	30
Figura 8 – Exemplo de arquivo em formato XML.....	34
Figura 9 – Stemming gerado pelo Sistema Aîuri.....	37
Figura 10 - Tela de apresentação do Sistema Aîuri.....	41
Figura 11 – Exemplo de uma matriz de confusão.....	46

LISTA DE TABELAS

Tabela 1 - Quantitativo de arquivos em cada classe para o idioma Português.....	34
Tabela 2 - Quantitativo de arquivos em cada classe para o idioma Inglês.....	35
Tabela 3 - Quantidade de arquivos utilizados em cada classe para treinamento e teste dos algoritmos – Idioma Português.....	42
Tabela 4 - Quantidade de arquivos utilizados em cada classe para treinamento e teste dos algoritmos – Idioma Inglês.....	42
Tabela 5 – Resultados para a classe Data Mining no idioma Português.....	46
Tabela 6 – Resultados para a classe Eficiência Energética no idioma Português.....	47
Tabela 7 – Resultados para a classe Energia no idioma Português.....	47
Tabela 8 – Resultados para a classe SIIF no idioma Português.....	48
Tabela 9 – Resultados para a classe Equipamento Elétrico no idioma Português....	48
Tabela 10 – Resultados para a classe Gerenciamento no idioma Português.....	49
Tabela 11 – Resultados para a classe Pesquisa e Desenvolvimento no idioma Português.....	50
Tabela 12 – Resultados para a classe Energia Eólica no idioma Português.....	50
Tabela 13 – Resultados para a classe Data Mining no idioma Inglês.....	51
Tabela 14 – Resultados para a classe Eficiência Energética no idioma Inglês.....	52
Tabela 15 – Resultados para a classe Energia no idioma Inglês.....	52
Tabela 16 – Resultados para a classe Educação no idioma Inglês.....	53
Tabela 17 – Resultados para a classe Energia Distribuída no idioma Inglês.....	53
Tabela 18 – Resultados para a classe Manutenção no idioma Inglês.....	54
Tabela 19 – Resultados para a classe Pesquisa e Desenvolvimento no idioma Inglês.....	55
Tabela 20 – Resultados para a classe Energia Eólica no idioma Inglês.....	55
Tabela 21 – Resultados para a classe Geração Distribuída no idioma Inglês.....	56
Tabela 22 – Resultados para a classe Mercado no idioma Inglês.....	56
Tabela 23 – Resultados para a classe Mudanças Climáticas no idioma Inglês.....	57
Tabela 24 – Resultados para a classe Qualidade de Energia no idioma Inglês.....	58
Tabela 25 – Resultados para a classe Renováveis no idioma Inglês.....	58
Tabela 26 – Resultados para a classe Sistemas de Potência no idioma Inglês.....	59
Tabela 27 – Resultados para a classe Termelétrica no idioma Inglês.....	59
Tabela 28 – Resultados para a classe Tecnologia da Informação no idioma Inglês..	60
Tabela 29 – Resultados para a classe Transmissão no idioma Inglês.....	60

LISTA DE ABREVIações E SIGLAS

IBM	International Business Machines Corporation
LEV	Laboratório de Energia dos Ventos
UFF	Universidade Federal Fluminense
PDF	Portable Document Format
RI	Recuperação de Informações
KDT	<i>Knowledge Discovery from Text</i>
BOW	<i>Bag Of Words</i>
PLN	Processamento de Linguagem Natural
ML	Machine Learning
SVM	Support Vector Machin
TXT	Texto Plano
XML	Extensible Markup Language
RNA	Rede Neural Artificial
AG	Algoritmos Genéticos
UFRJ	Universidade Federal do Rio de Janeiro
COPPE	Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia
NACAD	Núcleo de Atendimento em Computação de Alto Desempenho
TF-IDF	Term frequency-inverse document frequency
TF	Term frequency
DF	Document frequency

1 INTRODUÇÃO

A importância da informação para a sociedade atual é inquestionável. Vivemos atualmente na sociedade do conhecimento, onde a informação possui papel estratégico. Desde o início da utilização do computador para o tratamento de dados textuais, na década de 50, com as pesquisas de Hans Peter Luhn, e principalmente com a explosão do uso da Internet, em meados da década de 90, a informação disponível no formato digital, geralmente de acesso gratuito, cresce de forma espantosa, o que causa um problema “recente” em nossa sociedade: o da “inundação de informação”, que ocorre quando a pessoa recebe uma quantidade tal de informações, mesmo que relevantes, que não consegue tratá-las ou assimilá-las.

Anteriormente ao uso popular da Web, as buscas de informação eram restritas as bibliotecas, freqüentemente com a ajuda de um bibliotecário ou um sistema proprietário de recuperação de informações (KONCHADY, 2006). Pode-se afirmar que o grande problema até então era da “escassez de informações”, no tocante a sua localização.

A popularização do uso de computadores pessoais, a partir da década de 80, assim como o seu barateamento e aumento do seu poder de processamento e de armazenamento, aliado ao grande crescimento da Internet, tanto em quantidade de informações disponíveis como em relação a velocidade de acesso (*links*) cada vez mais rápida, a partir de meados da década de 90, permitiu a acumulação de grandes bases de dados textuais em computadores pessoais, geradas pelo próprio usuário ou obtidas através de diversos meios e em diversos formatos. Estes fatores geraram um grande acúmulo de informações, em grande parte textuais, nos computadores pessoais.

Segundo Konchady (2006) e Tan (1999), acima de 80% ou mais de informação de uma empresa estão armazenadas na forma de texto desestruturado. O manuseio deste imenso volume de informação é inviável sem ferramentas que permitam a localização de informação relevante quando nós precisamos dela.

Em entrevista a revista INFO EXAME (2006), Jean Paul Jacob, pesquisador brasileiro que atualmente trabalha no laboratório de IBM em Almaden, diz que passamos de seres onívoros, preocupados em obter alimento e energia para o corpo, para seres informívoros, que vivem de informação. Esta mesma afirmação, de

informação como algo fundamental, encontra-se em Le Codiac (2004): “A hipótese subjacente é que os seres humanos têm necessidade de informação da mesma forma que necessitam de alimento ou abrigo. A necessidade de informação tem então o status de uma necessidade física fundamental”.

Em uma alarmante reportagem publicada no jornal O GLOBO de outubro de 2006, a IBM prevê que a inundação de informação vai se tornar (quase) insuportável. Ainda nesta mesma reportagem, dados da IBM apontam que a informação disponível no mundo dobrará a cada 11 horas no ano de 2010. Em face deste problema torna-se imperiosa a aplicação de métodos computacionais modernos para o adequado tratamento das informações.

Grandes empresas possuem sistemas, desde a década de 70, para um adequado tratamento de suas informações. Porém, estas soluções tecnológicas são muito caras, complexas, podendo ser tecnologias proprietárias de alguns gigantes da informática, e demandam o uso de mão de mão-de-obra especializada para o seu correto uso, computadores com maior capacidade de processamento, redes lógicas rápidas, etc, tornando seus custos, tanto de implantação quanto de manutenção, muito elevados.

O método mais comum de organizar estes dados no formato digital para o usuário doméstico é baseado no uso de diretórios e nomes de arquivos apropriados. Esta forma de organização está se tornando ineficiente para gerenciar um grande número de informações em um computador pessoal. Até as pessoas mais organizadas podem achar dificuldade para lembrar centenas de nomes de diretórios e arquivos. Este processo não costuma produzir bons resultados quando se deseja recuperar as informações. Como já afirmara Davies, (1989 apud LOH, 2001) “Nada é mais frustrante do que saber que a resposta da questão está em algum arquivo no seu PC cujo local ou nome não se pode recuperar”.

As tecnologias atualmente disponíveis para o tratamento de informações permitem uma maior interatividade entre os usuários e as informações contidas nos documentos. Este impacto poderá ser visto futuramente nas bibliotecas. A biblioteca de hoje (um armazém de objetos passivos) fazendo uso das novas tecnologias emergentes, poderá fornecer ao usuário conexões desconhecidas, fazer associações e analogias, sugerir conceitos remotos ou novos, descobrir novos métodos, teorias, medidas... (LOH, 2001).

Atualmente coleta-se documentos, slides, arquivos de áudio e de vídeo, imagens, páginas Web, e-mails, códigos fonte, softwares e arquivos texto nos PCs.

Com o rápido crescimento da capacidade dos discos rígidos, torna-se fácil armazenar centenas de gigabytes de dados.

Pesquisadores, consultores e pequenas empresas armazenam uma quantidade enorme de informações, não somente textual, como em outros formatos (imagens estáticas, áudio, vídeo), gerados pelos próprios ou através, principalmente, da Internet, além da informação poder estar armazenada em outros meios, como o impresso. Neste cenário, o método de organização de informações que consiste na utilização de pastas e sub-pastas, além de nomes de arquivos relevantes ao conteúdo do mesmo, é um método ineficiente e custoso. Para tornar estas informações relevantes disponíveis aos seus reais usuários torna-se imperioso o emprego de métodos adequados para o correto tratamento da informação. Ferramentas baseadas em mineração de dados textuais (*text mining*) podem auxiliar na tarefa de manejo destes dados.

Text-mining, também conhecido como *text data mining* ou *knowledge discovery from textual databases* refere-se ao processo de extrair padrões e conhecimentos interessantes e não-triviais, de documentos não estruturados (TAN, 1999). Trata-se de uma tarefa mais complexa do que *data-mining*, já que envolve procedimentos com dados inerentemente desestruturados e pouco definidos.

1.1 FORMULAÇÃO DA SITUAÇÃO-PROBLEMA

A base de dados textuais do Laboratório de Energia dos Ventos da Universidade Federal Fluminense (LEV/UFF), é um caso típico da grande quantidade de informações textuais disponíveis nos PCs atuais.

Na tentativa de se realizar a organização desta massa documental, tentou-se a categorização manual utilizando um vocabulário controlado e uma codificação para cada classe descrita no vocabulário. Esta tentativa não obteve êxito por dois motivos: a grande quantidade de documentos e o seu rápido crescimento e; o vocabulário não fornecer as classes desejáveis para uma melhor organização.

Deste modo, a solução do problema seria a utilização de um sistema automático de baixo custo de implementação e operação, tendo em vista os recursos financeiros e humanos disponíveis no LEV.

A facilidade atual de se obter documentos digitais, assim como as crescentes áreas de atuação e de interesse do Laboratório, tornaram necessários a utilização de programas capazes de fornecer subsídios para facilitar o processo de categorização destes documentos.

A categorização facilita a identificação de documentos relevantes dentro de um mesmo sub-domínio. Para GALHO (2003), recuperar documentos em bases devidamente classificadas é sempre mais eficaz.

Sendo o ponto de partida do que se quer resolver, o problema identificado consiste em categorizar, visando um melhor gerenciamento de informações, uma grande base de dados, em torno de 22 gigabytes, sendo esta base formada em sua maioria por documentos no formato PDF. Detectou-se então, que a aplicação de ferramentas computacionais para gerir este material digital poderiam auxiliar na organização do acervo digital.

1.2 OBJETIVOS DA TESE

1.2.1 Objetivo Geral

A pergunta que se quer responder neste trabalho é se existem metodologias, algoritmos e softwares disponíveis para a solução eficiente e eficaz do problema de manuseio de uma grande quantidade de informações em empresas e instituições de pequeno porte, com recursos humanos e financeiros compatíveis com os disponíveis por estas empresas e instituições.

1.2.2 Objetivos Específicos

Os objetivos específicos desta dissertação são os seguintes:

- a) Levantamento Bibliográfico sobre o tema;
- b) Escolha da metodologia, dos algoritmos e do software a serem aplicados ao caso do Laboratório de Energia dos Ventos da UFF;
- c) Aplicação da metodologia, dos algoritmos e do software escolhidos no objetivo (b);
- d) avaliação dos resultados obtidos.

1.3 RELEVÂNCIA

A relevância do presente trabalho se dá pelos devida aos fatores citados a seguir:

a) Existe um problema real de manuseio de uma grande disponibilidade de informações, enfrentado por toda a sociedade, em especial empresas de pequeno porte, informações estas que chegam desordenadas, necessitando muito tempo para a localização de informações relevantes na solução de um dado problema;

b) Aplicação de técnicas computacionais modernas para a categorização de textos em um laboratório universitário de pesquisa em um caso real de uma grande quantidade de arquivos digitais textuais com nenhuma ou muito pouca organização;

c) Tratamento de arquivos, em português e em inglês, cada idioma com as suas peculiaridades em algumas etapas do processo de mineração de dados textuais para a categorização de textos em países onde a o idioma oficial é o português;

d) A importância de uma “melhor” recuperação de informação. – “Poupe o tempo do leitor”, a quarta lei da Biblioteconomia, do conjunto de cinco, que são as cinco leis fundamentais instituídas pelo pensador indiano Shiyali Ramamrita Ranganathan.

1.4 CAPÍTULOS SEGUINTES

Além do presente capítulo de INTRODUÇÃO, esta dissertação contém os seguintes capítulos:

Capítulo 2: Metodologia

Neste capítulo são descritos os passos utilizados neste trabalho para responder a pergunta básica do trabalho.

Capítulo 3: Revisão Bibliográfica

Neste capítulo foi feita uma pesquisa bibliográfica através da Internet e em bibliotecas, sobre a definição do que é informação, conceito de recuperação de informação, mineração de textos e categorização, principais temas abordados por esta dissertação.

Capítulo 4: Escolha da metodologia, dos algoritmos e do software

Após a conclusão da etapa 3, Revisão Bibliográfica, foram escolhidos a metodologia, os algoritmos e o software que aplicados ao caso real dos arquivos do Laboratório de Energia dos Ventos da UFF.

Capítulo 5: Aplicação ao caso do Laboratório de Energia dos Ventos

Neste capítulo abordou-se as características dos dados textuais recolhidos, além de parâmetros utilizados no processo de categorização dos *corpus* existentes no Laboratório de Energia dos Ventos.

Capítulo 6: Processamento dos dados

Neste capítulo aplicou-se os categorizados de texto ao caso real da base textual do Laboratório de Energia dos Ventos. Constatou-se os resultados desta aplicação.

Capítulo 7: Avaliação dos Resultados

Neste capítulo foi descrito os resultados da pesquisa e analisado o desempenho dos categorizadores nos *corpus* utilizados.

Capítulo 8: Conclusão e Sugestões de Novos Trabalhos

Aborda os resultados da pesquisa, além de sugestões para trabalhos futuros.

Além dos capítulos, no fim desta dissertação há a listagem da bibliografia utilizada, e dois apêndices, cada um contendo a listagem dos arquivos utilizados na formatação da base textual, separados por idioma.

2 METODOLOGIA

As etapas da metodologia empregada para elaboração deste trabalho são descritas a seguir.

Etapa 1) Levantamento Bibliográfico

Identificou-se o material relevante sobre a temática proposta na dissertação. Foram feitas buscas em artigos de periódicos, páginas *Web* e livros. Buscou-se identificar as metodologias de categorização, bem como os softwares disponíveis e os algoritmos de categorização mais abordados na literatura;

Etapa 2) Escolha do conjunto de Metodologia, algoritmos de categorização de textos e software a ser utilizado.

Embasado pelo levantamento bibliográfico, escolheu-se o conjunto composto pela metodologia viável para o tipo de *corpus* a ser tratado, pelos algoritmos viáveis de serem utilizados e pelo do software adequado.

Etapa 3) Aplicação ao caso do Laboratório de Energia dos Ventos

Nesta etapa executou-se os passos da metodologia escolhida e aplicou-se o sistema escolhido nos *corpus* a serem analisados.

Etapa 4) Avaliação dos Resultados

Nesta etapa consta os resultados obtidos com a dissertação, “respondendo” a questão principal norteadora da tese, que é se a aplicação de técnicas de categorização auxiliam o usuário final a localizar a informação desejada. Esta etapa foram usados os parâmetros de avaliação a precisão, a revocação e a média F e a matriz de confusão.

Etapa 5) Conclusão e Recomendações para novos trabalhos

Nesta etapa descreve-se os resultados obtidos com a categorização, além de sugestões para trabalhos futuros nesta área.

Etapa 6) Referências Bibliográficas

Nesta etapa listou-se a bibliografia referenciada na dissertação.

3 REVISÃO BIBLIOGRÁFICA

Para a elaboração da revisão bibliográfica buscou-se material em diversas fontes, tais como, o portal de periódicos da CAPES, em especial os portais do IEEE e ACM, o Banco Digital de Teses e Dissertações de diversas Universidades e Bibliotecas, além de material disponibilizado em sites de Universidades Brasileiras e do exterior. Foram também utilizados motores de buscas com termos das temáticas envolvidas nesta dissertação.

3.1 INFORMAÇÃO

Informação, se definida de forma simples, é um conjunto de dados ordenados com alguma significação. Por exemplo, Rio de Janeiro e 28°C são dados, porém se o colocarmos com algum sentido: No Rio de Janeiro fez 28°C, estes dados transformam-se em uma informação.

Meadows (2003) traça a diferença entre “dados” e “informações”. Para o referido autor, os dados brutos são aqueles obtidos por medições e observações diretas, e provavelmente vão continuar a crescer de volume no futuro indefinido. A informação, para ele científica, é obtida através da análise e discussão destes dados.

O conceito de informação na sociedade atual, segundo FERNEDA (2003), é a que permite sua operacionalização ao através do computador ou outros dispositivos digitais. Diante deste conceito, torna-se evidente a necessidade de seu tratamento usando ferramentas computacionais, visto que já é inviável a organização de uma grande massa de documentos sem auxílio computacional.

Segundo McGarry (1999, apud FERNEDA, 2003,) a palavra “informação” tornou-se popular logo após a invenção da imprensa no século XV, quando normalmente se utilizava uma palavra em latim para expressar uma nova idéia ou conceito. A raiz do termo vem de *formatio* e *forma*, ambos transmitindo a idéia de “moldar algo” ou dar “forma a” algo indeterminado.

A idéia de um volume de informações gerado sendo maior do que a capacidade humana de utilizá-las não é nova. A invenção da imprensa causou um

aumento significativo no volume de material textual disponível, o que causou um grande impacto na difusão das informações.

Encontramos em Meadows (2003) um autor que em 1613, portanto depois da invenção da imprensa na Europa, no século XV, já tinha salientado que “um dos males destes tempos é a multiplicidade de livros; eles, de fato, sobrecarregam de tal modo a gente que não conseguimos digerir a abundância de matéria inútil que, todos os dias, é gerada e despejada no mundo”.

A explosão informacional pós 2ª Guerra Mundial, gerou a necessidade de busca de soluções do problema advindo da avalanche de informações. Em 1945 Vannevar Bush aborda o problema do gerenciamento da informação e propõe como solução uma máquina, denominada Memex, que agregava as mais modernas tecnologias de informação existentes na época.

A explosão informacional pós-guerra, além do grande desenvolvimento tecnológico, gerou uma grande quantidade de informações que necessitavam de tratamento. Em 1949, Palmer já havia diagnosticado a necessidade de sistemas capazes de superar a produção de conhecimentos que crescia diariamente. A busca de soluções para o problema da explosão informação tem início na década de 50, tendo destaque o pesquisador da IBM Hans Peter Luhn. Na década de 70 apareceram os Editores de texto, auxiliando o crescimento da massa informacional. Na década de 90, com o uso da Internet difundido, aumentou-se ainda mais a quantidade de informações e a sua disponibilização.

Realizando a associação entre dados e informação, Hayes (1986, apud FERNEDA, 2003) apresenta a seguinte definição: “Informação é uma propriedade dos dados resultante de/ou de produzida por um processo realizado sobre os dados. O processo pode ser simplesmente a transmissão dos dados (cujo casos são aplicáveis a definição e medida utilizadas na teoria da comunicação); pode ser a seleção de dados; pode ser a organização de dados; pode ser a análise de dados”.

Segundo Marcondes (2003), atualmente temos a migração acelerada de registros da cultura humana para suporte eletrônico na Internet; número crescente de registros já criados diretamente no meio eletrônico; diversidade de tópicos de interesse, de áreas de conhecimento, de idiomas, de públicos e; diversidade de estruturas de suporte de conhecimento, incluindo textos, hipertextos, imagens estáticas e dinâmicas, hipermídias... O grande problema é identificar e permitir acesso às informações. Para Marcondes (2003) “Se a informação está disponível na Internet

mas não é encontrada, o conhecimento não é realizado”.

Muitas vezes confunde-se informação com documento. Em Le codiac (2004) encontra-se a definição de documento como sendo o termo genérico que designa os objetos portadores de informação. O documento vem a ser somente o suporte (material ou digital) da informação, podendo ser de diversos tipos: textos, sons, vídeos, imagens estáticas, dentre outros.

3.2 RECUPERAÇÃO DE INFORMAÇÕES

A área de Recuperação de Informações (RI) destina-se a auxiliar as pessoas a encontrar documentos que tenham informações relevantes. Para o assunto que está sendo tratado, entretanto, é necessário examinar os documentos resultantes da busca para encontrar a informação desejada. A dificuldade vem do fato de que uma busca em uma ferramenta computacional pode trazer nenhum documento como resultado ou o inverso, uma quantidade muito grande de documentos, causando a chamada “sobrecarga de informação” (*information overload*), que acontece quando o usuário tem muita informação disponível, mas não tem condições de tratá-la ou encontrar o que realmente interessa. O processo de Recuperação de Informação consiste em identificar, no conjunto de documentos de um sistema, quais atendem à necessidade de informação do usuário.

Os sistemas de recuperação de informação devem representar o conteúdo dos documentos do corpus e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente à sua necessidade de informação, formalizada através da expressão de busca.

Os primeiros sistemas de recuperação de informação baseavam-se na contagem de frequência das palavras do texto e na eliminação de palavras reconhecidamente de pouca relevância. Salton e Mc Gill (1983 apud FERNEDA, 2003) abordam a aplicação do processamento da linguagem natural e da lógica *fuzzy* na recuperação de informação, apontando a direção de futuras pesquisas para a Inteligência Artificial.

A técnica mais básica e mais usada para Descoberta de Conhecimento em textos (*Knowledge Discovery from Text – KDT*) é a recuperação de informação (RI), que se limita a encontrar documentos ou textos onde informações relevantes possam

estar. O termo *Knowledge Discovery from Text* foi utilizado pela primeira vez por Feldman e Dagan (LOH, 2001) para designar o processo de encontrar algo interessante em coleções de textos (artigos, e-mails, páginas Web, dentre outros).

3.3 TÉCNICAS DE MINERAÇÃO DE TEXTOS

3.3.1 Introdução e Considerações Gerais

Conforme afirma Ebecken, Lopes e Costa (2003), todos os tipos de textos que compõem o dia-a-dia de empresas e pessoas são produzidos e armazenados em meios eletrônicos. Inúmeras novas páginas contendo textos são lançadas diariamente na web. Outros tipos de documentos como: relatórios de acompanhamento, atas de reuniões e históricos pessoais são periodicamente gerados e atualizados. Entretanto, até pouco tempo atrás, essas informações em formato de textos não eram usadas para significar algum tipo de vantagem competitiva, ou mesmo como suporte à tomada de decisões, ou ainda como indicador de sucesso ou fracasso. Com o advento da “mineração de textos”, a extração de informações em textos passou a ser possível e o imenso e crescente mundo dos textos está começando a ser explorado.

Mineração de textos (*Text mining*) tem sido crescentemente empregado para denotar todas as tarefas de análise de grandes quantidades de texto. A origem do que nós chamamos *text-mining*, também conhecido como *Knowledge Discovery from Text* e *Text Data Mining*, vem da área de recuperação de informação (WEISS, 2005). A classificação de documentos é similar em muitas formas a indexação de documentos que foi estudada extensivamente no fim dos anos 50 e nos anos 60. Agrupamento de documentos e medidas de similaridade entre documentos também são tópicos antigos. A representação dos documentos como BOW (*Bag Of Words*), tornaram-se popular na década de 70.

3.3.2 Abordagens dos dados utilizados em Mineração de Texto

Segundo Ebecken, Lopes e Costa (2003), há duas formas principais de abordagem dos dados textuais. A análise semântica, baseada na funcionalidade dos termos nos textos e a análise estatística baseada em freqüência.

3.3.2.1 Análise semântica dos dados

Depois de separar o texto em *tokens* (geralmente termos) e sentenças, o próximo passo depende do que será feito com o texto. Se não há necessidade de análise lingüística, procede-se diretamente com a geração de características, em que as características serão obtidas diretamente dos *tokens*.

Porém, se a meta é mais específica, como o reconhecimento do nome de pessoas, lugares e organizações, é normalmente desejável executar análises lingüísticas adicionais no texto para se extrair características mais sofisticadas.

3.3.2.2 Análise estatística baseada em freqüência

Nesta análise a importância dos termos é dada basicamente pelo número de vezes que elas aparecem nos textos. Este tipo de estratégia independe do idioma a ser abordado e foi a utilizada nesta dissertação.

3.3.3 Etapas do processo de Mineração de Textos

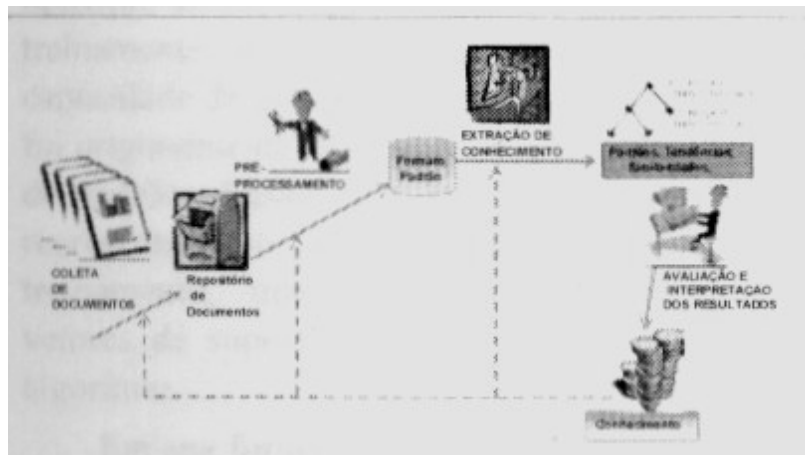
O processo de mineração de textos pode ser dividido, segundo Gonçalves e Resende (20-?), em quatro grandes etapas:

- Coleta de Documentos: Etapa onde documentos relacionados com a tarefa são recolhidos;
- Pré-processamento: Nesta etapa são realizadas ações sobre o material recolhido com o objetivo de prepará-los para a extração de conhecimento. Geralmente têm-se o resultado a padronização dos documentos em um formato atributo-valor;

- Extração de Conhecimento: Nesta etapa são utilizados os alguns algoritmos de aprendizado com o objetivo de extrair o conhecimento na forma de regras de associação, relações, segmentação, classificação de textos, entre outros.
- Avaliação e Interpretação dos Resultados: Nesta etapa os resultados obtidos são analisados, filtrados e selecionados para que o usuário possa ter um melhor entendimento dos textos coletados.

As etapas podem ser observadas na figura 1.

Figura 1: Principais etapas do processo de Mineração de Textos



FONTE: Gonçalves e Rezende [20-?]

3.3.4 Principais técnicas de mineração de textos

A Mineração de Textos é um campo multidisciplinar, envolvendo recuperação de informações, análise de texto, extração de informação, agrupamento, categorização, visualização, banco de dados, aprendizado de máquina e mineração de dados (TAN, 1999). Embora o campo seja relativamente novo, ele é dependente de um amplo corpo de pesquisas em Recuperação de Informação (RI) e Processamento de Linguagem Natural (PLN). O emprego de métodos estatísticos, juntamente com algum dicionário de definições, é o método mais comum para a construção de ferramentas de mineração de textos.

Não existe um conjunto padrão de aplicações de mineração de textos. As aplicações mais comuns são:

a) Recuperação de informação, que se destina a encontrar documentos que contenham informações relevantes em relação a busca do usuário. As técnicas de RI podem ajudar apresentando documentos com visão geral das informações ou assuntos (RI tradicional) ou apresentando partes de documentos com detalhes de informações (recuperação por passagens). Também as ferramentas de RI por filtragem contribuem garimpando documentos interessantes para o usuário, sem que este precise formular consultas (LOH, 2001);

b) Extração de informações, que visa encontrar termos específicos nos textos;

c) *Clustering*, que é o agrupamento de texto de acordo com semelhanças ou similaridades no conteúdo, não havendo classes pré-definidas. A técnica de agrupamento é diferente da categorização, pois o agrupamento visa criar as classes através da organização dos elementos, enquanto a categorização procura alocar os documentos em categorias pré-determinadas. Para Loh (2001), o agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes;

d) Categorização de texto, que é classificar um documento segundo categorias (tópicos ou temas) preexistentes de documentos. A categorização difere-se do agrupamento, pois na categorização se conhece previamente as características de cada classe, enquanto que no *clustering* não se conhece previamente as características de cada classe; e

e) Sumarização, que consiste em resumir textos, tentando identificar e extrair as idéias mais importantes.

Dentre as principais técnicas de mineração de texto, optou-se pela técnica de categorização de textos, pois era a que melhor atendia as necessidades do Laboratório.

3.4 CATEGORIZAÇÃO DE TEXTO

Segundo visto no tópico “d” anterior, a tarefa de categorização visa a alocação de documentos em tópicos ou temas pré-definidas. Antes da possibilidade do uso de

computadores para a tarefa, a categorização era feita de forma manual, geralmente em centros de documentação. Porém, com o volume atual de informações disponíveis, este processo torna-se impraticável.

3.4.1 Categorização Manual

A categorização manual categorizar os documentos em uma ou mais classes pré-determinadas. Necessita do uso de um vocabulário controlado, além da análise técnica do documento. Geralmente realizada em centros de documentação. É um processo demorado e caro. Por exemplo, MEDLINE (*National Library of Medicine*) usa dois milhões de dólares por ano para indexar manualmente os artigos de revista. Outro exemplo é o site de buscas Yahoo, que possui mais de 200 experts para rotular manualmente ou categorizar os sites, que recebe centenas de páginas diariamente. (SHALABI, KANAAN e GHARAIBEH, 2004).

3.4.2 Categorização Automática

A Categorização de textos, também conhecida como classificação, é uma tarefa de aprendizado supervisionado, definido como a atribuição de “labels”/categorias (pré-definidas) para novos documentos baseados na semelhança do conjunto de treinamento dos documentos categorizados. (LOPES, 2004).

A Categorização Automática de Texto atualmente é uma disciplina obtida com o cruzamento de ML (*Machine Learning*) e RI, assim como compartilha um número de tarefas com outras técnicas tais como *information extraction* e sumarização.

A categorização automática de textos possui diversas vantagens em relação a categorização manual. A categorização feita manualmente por pessoas, segundo Wives (1999, p. 12) acarreta problemas de atraso (já que há um limite no número de informações que podem ser indexadas diariamente por um ser humano) ou de indexação imprecisa (onde a pessoa que indexa pode não categorizar corretamente a informação, colocando-a em uma categoria diferente da categoria que a informação realmente pertence).

São utilizados métodos que identificam os conceitos no texto e efetuam a categorização de fato. Esses métodos podem classificar os documentos em nenhuma,

uma ou mais categorias existentes. Quando um método efetua a categorização de textos em apenas uma categoria, diz-se que este método é de classificação binária. E, quando os textos podem ser classificados em mais de uma categoria, diz-se que foi aplicado o método de categorização graduada, podendo definir o grau de pertinência do documento a cada uma das categorias para as quais ele foi classificado (LEWIS apud RIZZI, 2000).

O estudo de Categorização de Textos utilizando computadores data do início de 1960, porém somente no início da década de 90 tornou-se um subcampo maior na disciplina de sistemas de informação, graças o aumento do interesse por aplicativos e a disponibilidade de hardware mais poderosos.

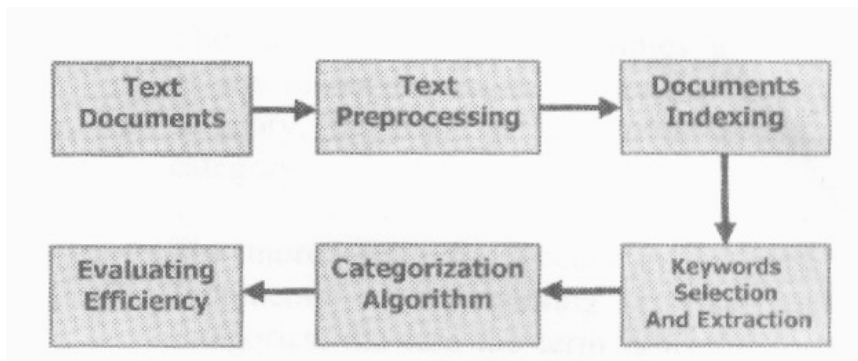
Até a década de 80 a abordagem mais popular para a categorização de textos, pelo menos na comunidade “operacional” (aplicações do mundo real) era a Engenharia do Conhecimento (*Knowledge Engineering* - KE), que consiste na definição manual de um conjunto de regras, codificadas por um especialista, de como classificar os documentos sobre as categorias fornecidas. Na década de 90 esta abordagem foi perdendo popularidade (especialmente na comunidade de pesquisa) para a abordagem de aprendizado de máquina (ML), onde um processo geral indutivo constrói automaticamente um classificador que aprende de um conjunto de documentos pré-classificados as características da categoria. (SEBASTIANI, 2002).

A vantagem da abordagem do aprendizado de máquina são a acurácia comparável com a alcançável por especialistas humanos, e a economia da mão de obra do especialista, assim como nenhuma intervenção de um engenheiro de conhecimento ou especialista do domínio é necessária para a construção do classificador ou encaminhar a um conjunto diferente de categorias.

As categorias são escolhidas para corresponder aos tópicos ou temas dos documentos. O principal objetivo da categorização de textos é a organização automática. Alguns sistemas de categorização (ou categorizadores) retornam uma única categoria para documento, enquanto outros retornam categorias múltiplas. Em ambos os casos, um categorizador pode retornar nenhuma categoria ou algumas categorias com confiabilidade muito baixa. Nestes casos, o documento é normalmente associado a uma categoria rotulada como “desconhecida” ,para posterior classificação manual (LOPES, 2004).

O esquema do processo de categorização pode ser visto na figura abaixo:

Figura 2 - Processo de categorização



FONTE: AL SHALABI, KANAAN e GHARAIBEH (2004)

Atualmente se emprega categorização de textos em muitos contextos, tais como, a indexação automática para sistemas booleanos de recuperação de informação, organização de documentos, filtragem de textos, *word sense disambiguation*, e categorização hierárquica de páginas *Web* (SEBASTIANI, 2002).

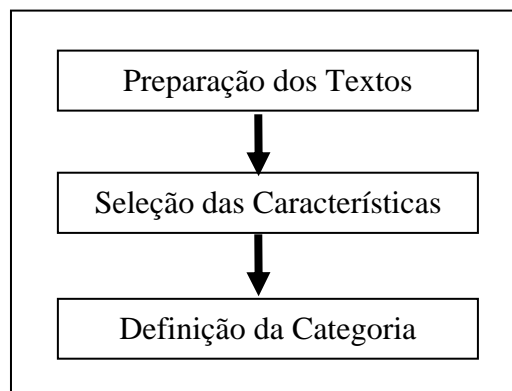
Para GALHO (2003), a técnica pode ser usada como uma forma de organizar os documentos tanto para a recuperação quanto para a armazenagem. A classificação facilita o processo de busca, pois ao invés de se realizar a busca em todo o *corpus* busca-se somente nas categorias de interesse.

O processo de categorização automática tradicional precisa de duas etapas: uma etapa de aprendizado, onde as classes são identificadas e caracterizadas, e uma etapa de classificação propriamente dita, onde os elementos são identificados (classificados) de acordo com as classes existentes (WIVES, 1999).

A etapa de aprendizado da categorização se realiza geralmente em três etapas: preparação de textos, seleção das características e definição das categorias.

A figura 3 mostra o esquema desta etapa.

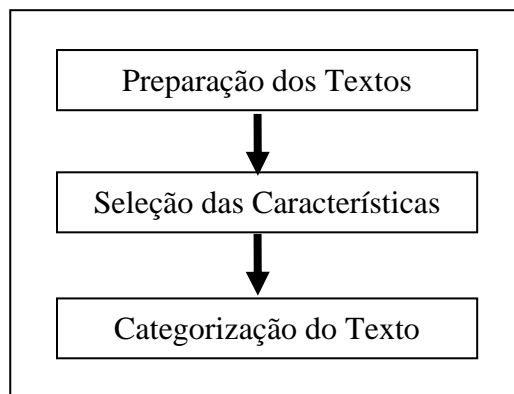
Figura 3: Esquema da etapa de aprendizado



Fonte: GALHO, 2003

Na etapa de classificação o novo texto também é preparado e suas características são selecionadas. A categorização ocorre através da comparação entre o índice da categoria e o índice do texto. A figura 4 mostra o esquema desta etapa.

Figura 4: Esquema da Etapa de classificação



Fonte: GALHO, 2003.

Os métodos automáticos de indexação geralmente utilizam “filtros” para eliminar palavras de pouca significação, além de normalizar os termos reduzindo-os a seus radicais.

Alguns métodos de categorização são descritos a seguir:

3.5 MÉTODOS DE CATEGORIZAÇÃO AUTOMÁTICA USADOS EM APRENDIZADO DE MÁQUINA

3.5.1 - Espaço vetorial e similaridade por co-seno

O modelo de indexação por espaço vetorial para RI considera o documento representado por um vetor de termos e seus respectivos pesos associados. O vetor possui a forma {(termo 1, peso 1), (termo 2, peso 2)... (termo n. peso n)}, onde n representa a n-ésima dupla termo-peso representante do documento. Calcula da similaridade ou diferença entre dois vetores através da medida do co-seno entre eles. O ângulo entre dois vetores pode ser apresentado no espaço euclidiano.

Este processo pode ser utilizado para categorização automática de documentos textuais visto que documentos e categorias podem ser representados por

vetores. Através do cálculo do co-seno do ângulo entre estes vetores, é possível definir a similaridade entre eles. Quanto menor for o ângulo entre dois vetores, maior a similaridade entre os documentos que eles representam. Salton e Buckley (1987) propuseram uma fórmula para cálculo da similaridade por co-seno entre vetores normalizados.

3.5.2 - K-NN

A classificação do k vizinho mais próximo (*K nearest neighbour* - K-NN), é uma conhecida abordagem estatística, intensivamente estudada em reconhecimento de padrões há mais de quatro décadas. K-NN tem sido aplicado para categorização de textos desde os estágios iniciais da pesquisa em categorização (YANG e LIU, 1999).

O método K-NN avalia a classe de um texto candidato pelas classes associadas aos textos mais semelhantes a ele. Não há descritores para as classes, mas somente casos passados. Assim, a função de similaridade avalia a semelhança entre o conteúdo dos textos tomando palavras como características (LOH, 2001).

O algoritmo K-NN é muito simples: dado o documento teste, o sistema procura os k vizinhos mais próximos dentre os documentos de treino, e usa as categorias dos k vizinhos para pesar as categorias candidatas. A contagem da similaridade de cada vizinho documento para o documento teste é usada como o peso da categoria do documento vizinho. Se muitos vizinhos próximos compartilham a categoria, então o peso por-vizinho de cada categoria é adicionada juntamente, e a consequente soma ponderada é usada como a contagem da probabilidade dessa categoria em relação ao documento teste. Sorteando as contagens das categorias candidatas, uma lista rankeada é obtida para o documento teste. (YANG e LIU, 1999).

Yang (2000) diz que o desempenho do k-NN é dependente do valor de k e que, para a categorização de textos, avaliações demonstraram que um bom desempenho é alcançado quando o valor de K for grande, exemplificando com valores entre 30 a 200 K vizinhos.

Os resultados da aplicação do método k-NN são bons quando houver uma grande base de casos de treinamento. O uso do K-NN na classificação de notícias alcançou de 70 % a 80 % de acertos.

3.5.3 - Método Rocchio

O algoritmo de Rocchio (1971) foi inicialmente definido para o processo de relevância e feedback em RI, mas também é aplicado à categorização. Ele primeiramente constrói um vetor (vetor protótipo), para representar cada categoria; no caso da categorização de textos, um vetor de conceitos da categoria. Para efetuar a categorização de textos, é feito um cálculo de similaridade ou de distância entre o vetor de conceitos do texto e o vetor protótipo da categoria. Considerando o grau de similaridade dentro de um limiar pré-definido, o texto será classificado ou não para a categoria testada.

O Método Rocchio utiliza um vetor protótipo para descrever uma classe. Assim, palavras podem ser usados como elementos deste vetor e, portanto, pode-se avaliar a força de ambos como representantes do conteúdo dos textos e como descritores de classes (LOH, 2001). Apesar de não ser considerado o melhor método, é o mais simples.

O algoritmo de Rocchio não é o melhor, mas é simples de ser implementado. Resultados obtidos com a categorização de textos médicos foram inferiores a 50%.

3.5.4 - Redes Neurais

A técnica de rede neuronal (também chamado de método conexionista) é uma técnica que tem sido intensivamente estudada em Inteligência Artificial. Ela pode ser aplicada à categorização de textos, pois ela é comumente usada para classificação de padrões. Logo, pode mapear um texto de entrada em alguma categoria na saída.

O classificador de textos redes neurais é uma rede de unidades, onde a unidade de entrada representa termos, a unidade de saída representa a categoria ou categorias de interesse, e os pesos das unidades de pontes de conexão, representam as relações de dependência. Para classificar um documento teste d_j , os pesos do termo w_{kj} são carregados para as unidades de entrada; a ativação destas unidades e propagada através da rede, e o valor da unidade de saída determina a decisão de categorização. Um forma típica de treinar as redes neurais é a *backpropagation*, qual os pesos dos termos do documento de treino é carregado para as unidades de entrada, e se um erro de classificação o erro é “propagado de volta” para trocar os parâmetros de rede e eliminar ou minimizar o erro (SEBASTIANI, 2002).

Em Moens (2000 apud GALHO, 2003), a rede neuronal é definida como uma cadeia de neurônios artificiais, conectados em camadas. Cada um desses neurônios possui um peso de ativação. Para treinamento da rede, ainda são utilizadas as constantes bias e taxa e aprendizado. A cada exemplo inserido na rede para treinamento, os pesos das conexões são recalculados e atualizados. Diz-se que a rede está treinada quando os valores desses pesos permanecem constantes, ou após alcançar um critério pré-definido.

A rede neuronal pode também ser generalizada (intervalo de aceitação nos resultados - capacidade de se adaptar a novas situações e suportar ruídos), sendo necessário a utilização de bons casos de treinamento, para que se possa definir um intervalo de aceitação que não seja muito abrangente e nem muito fechado.

Depois de treinada e generalizada, a rede neuronal está pronta para receber novos textos e categorizá-los. Os exemplos de estudos, citados em Moens (2000, apud GALHO, 2003), apontam uma margem de 40-50% e acerto na análise de novos casos comparando com a categorização feita por especialistas.

3.5.5 - Modelo Fuzzy

A categorização de documentos trata da representação de textos através de termos e, por isso, enfrenta situações ambíguas para definir a relevância dos termos no que se refere à representação de uma categoria.

O uso da lógica fuzzy para a categorização de textos vem ao encontro da solução do problema de ambigüidade, pois a mesma se propõe a tratar situações imprecisas, oferecendo melhores resultados através do cálculo da pertinência de um elemento a um conjunto.

Os conjuntos que representam os documentos são compostos pelas duplas {termo, peso}, sendo o peso um valor fuzzy definido entre 0 e 1. Este valor indica a importância do termo, quanto mais próximo do valor um, mais relevante é o termo.

A partir da atribuição da relevância dos termos em relação ao documento, os sistemas fuzzy baseiam-se na idéia de similaridade, permitindo que os resultados ofereçam não apenas classificações exatas de um documento com relação a uma categoria, mas também categorizações parciais, sendo atribuída a cada classe um grau de pertinência ou de relevância com relação ao documento analisado.

3.5.6 - Árvores de Decisão

Este modelo induz a classificação por regras de decisão em árvores, a partir de exemplos de treinamento, e pode classificar novos casos, ou seja, casos ainda não vistos.

Cada regra de decisão está associada a uma categoria. As regras possuem a estrutura “se-então” e podem ser avaliadas como verdadeiro ou falso.

A classificação por árvores se dá pela divisão de um conjunto de regras de decisão. As árvores são constituídas por nodos e ramos. Cada nodo, exceto nodos terminais que são as categorias, representam um teste de decisão, e os ramos em subárvores representam cada possível resultado destes testes.

A construção de regras é feita a partir de um conjunto de exemplos positivos e um conjunto de exemplos negativos de uma dada categoria. As árvores de decisão são bem aplicadas quando existem dependências entre as categorias, o que é muito comum em tratamento de texto.

3.5.7 - Support Vector Machines

Support Vector Machines (SVM) é uma abordagem relativamente nova de aprendizado introduzida por Vapnik em 1995 para resolver problemas de reconhecimento de padrões de duas classes (YANG e LIU, 1999). O método é definido sobre o espaço de vetores onde o problema é encontrar a superfície de decisão que melhor separa os pontos de dados em duas classes.

É baseado no princípio de minimização de risco estrutural (Structural Risk Minimization Principle). O SRM minimiza o erro de generalização, que é a faixa de erro do algoritmo de aprendizado nos dados de teste. Essa característica permite ao algoritmo SVM uma maior capacidade de generalização. (Gonçalves e Rezende, 20-?)

Em sua forma linear, SVM gera um hiperplano que separa um conjunto de amostras positivas de um conjunto de amostras negativas. Em termos geométricos este método pode ser visto como uma tentativa de busca da melhor superfície σ_i , no conjunto de todas as superfícies $\sigma_1 \sigma_2 \dots \sigma_n$ no espaço r-dimensional que separa os exemplos de treinamento positivos dos negativos (superfície de decisão

(SEASTIANI, 2002). A melhor superfície σ_i separa os exemplos positivos dos negativos pela mais larga margem possível.

3.5.8 - Classificador Bayesiano

Utiliza uma metodologia parecida com a Rocchio, porém baseada em cálculos probabilísticos. A probabilidade do elemento pertencer a uma classe é avaliada pela comparação entre os vetores representativos. Neste caso, o centróide (ou vetor protótipo) da classe define os termos que provavelmente aparecem num texto da classe. O peso associado é a probabilidade de o termo aparecer em documentos da classe. Quanto mais termos da classe o texto contiver, maior a probabilidade de ele pertencer àquela classe. O método Bayesiano assume que não há dependência entre os termos, isto é, a probabilidade de um termo não é condicionada por outro. Este é o método mais simples entre os avaliados, exigindo pouca computação.

O funcionamento do método bayes é da seguinte forma: um conjunto de características é selecionado para representar uma categoria. Quando é feita a análise de um novo caso, é calculada a probabilidade das suas características estarem relacionadas com as características das classes armazenadas na base de casos. Para todas as classes são avaliadas as probabilidades com o texto, e aquelas com maior valor de probabilidade serão atribuídas a ele.

3.5.9 - Classificador Linear

Considere o problema de se distinguir entre duas classes. O método de contagem geral atribui a contagem positiva para a predição das classes positivas e a e a contagem negativa para a predição das classes negativas. A figura 5, abaixo, mostra um exemplo do uso de um conjunto de pesos para determinar a contagem para o documento. Para todas as palavras que ocorrem no documento, o modelo encontra os pesos correspondentes delas. Estes pesos são somados para determinar o contagem do documento (WEISS, 2004).

Figura 5 - Exemplo do uso de um conjunto de pesos para determinar a contagem para o documento

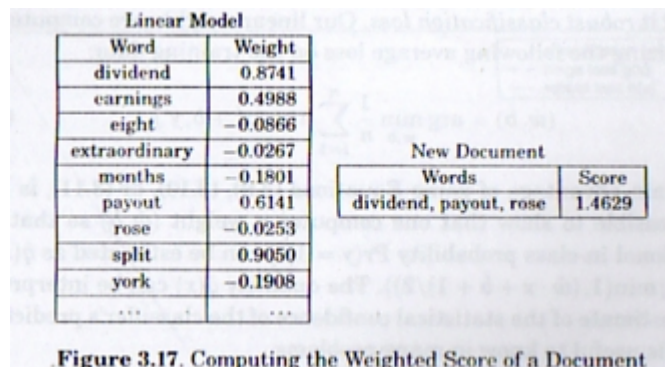


Figure 3.17. Computing the Weighted Score of a Document

Fonte: WEISS, 2004)

Métodos de contagem linear são abordagens clássicas para resolver problemas de predição. Geometricamente, o método pode ser descrito como produtor de uma linha ou hiperplano.

3.6 METODOLOGIA DO PROCESSO DE CATEGORIZAÇÃO

Como os categorizadores implementados no sistema Añuri trabalham com a análise estatísticas dos termos, aplicou-se os seguintes passos para a sua utilização.

3.6.1 Coleta dos textos

A primeira etapa do processo de mineração de dados textuais é a obtenção da base e a sua preparação para ser utilizada pelo algoritmo. Segundo Ebecken, Lopes e Costa (2003,), a preparação dos dados é a primeira etapa do processo de descoberta de conhecimento em textos. Esta etapa envolve a seleção dos documentos que constituirão a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo dos textos, ou seja, toda a informação que não refletir nenhuma idéia considerada importante poderá ser desprezada.

3.6.2 Pré-processamento

Após a coleta dos dados inicia-se o pré-processamento. Segundo Ebecken, Lopes e Costa (2003) o pré-processamento atua preparando o conjunto de dados textuais para as fases posteriores de execução das tarefas de processamento dos dados e análise de resultados obtidos. A figura 6 é um esquema da etapa de pré-processamento.

Figura 6 - Etapa de Pré-processamento



(fonte: Ebecken, Lopes, Costa, 2003)

3.6.2.1 Case Folding

O case folding é a conversão dos caracteres dos textos em um mesmo tipo de letra, geralmente caixa baixa. Este procedimento é necessário, pois o sistema é *case sensitive*, ou seja, diferencia letras em caixa alta de letras em caixa baixa.

3.6.2.2 Identificação de Termos

O processo de identificação de termos (“tokenização”), visa “quebrar” o fluxo de caracteres em palavras, ou mais precisamente, *tokens* (WEIS, 2004). Os caracteres de espaço, tabulação, e nova linha são sempre delimitadores e não são contados como *tokens*. Eles são coletivamente frequentemente chamados de *white space*. Os caracteres () <>!?” são sempre delimitadores e podem ser *tokens*. Os caracteres.,:;-‘ podem ou não ser delimitadores, dependendo do ambiente.

O ponto final, a vírgula e os dois pontos entre números não podem ser considerados delimitadores, mas parte do número. O ponto final pode ser parte de uma abreviação. O problema é detectar quando o ponto é o fim de uma sentença ou

não. Para a proposta da tokenização, é ,provavelmente, melhor tratar qualquer ponto ambíguo como uma palavra delimitadora e também como um *token*.

A tokenização é o processo que consiste identificar *tokens*, ou palavras, em um texto. Para Santos (2002) uma regra prática para identificar palavras, baseada em noções puramente gráficas, sugere que estas são definidas como "uma string com caracteres alfanuméricos contíguos sem espaços, podendo também incluir hífens e apóstrofes, mas nenhuma outra marca de pontuação".

Alguns problemas específicos são discutidos a seguir:

a) Ambigüidade do Ponto Final ("."): as palavras podem ter anexada a elas uma pontuação representada por vírgula, ponto-e-vírgula e ponto final. À primeira vista, apresenta-se fácil o reconhecimento da pontuação. Porém o caso de ponto final é problemático. Apesar de que muitas vezes um ponto marca o fim de uma sentença, às vezes um ponto faz parte de abreviaturas tais como etc. ou Calif. Presume-se que esses pontos das abreviaturas fazem parte da palavra com a qual se apresenta. Por exemplo: distingue-se Wash. (que é a abreviatura do estado de Washington) da forma capitalizada do verbo wash. Percebe-se também que quando uma abreviatura como etc. aparece no fim de uma sentença, apresenta-se apenas um ponto com ambas as funções de abreviatura e final de frase ao mesmo tempo.

b) Ocorrência de apóstrofes ('): Nesse caso é difícil saber como resgatar informações do tipo I'll ou isn't. Conforme definição anterior, apresenta-se como uma "palavra gráfica", mas há uma forte intuição que realmente tem-se duas palavras como as contrações de I will e is not. Assim, alguns sistemas separam tais contrações em duas palavras enquanto outros não o fazem.

c) Hifenização: Muitas expressões escritas de forma hifenizada são claramente tratadas como uma palavra única, como por exemplo e-mail, co-operate ou A-1-plus. Existem casos também como non-lawyer, pro-Arab e so-called aonde os hífens são léxicos. Eles são comumente inseridos antes ou após pequenas palavras de caráter formativo, algumas vezes com o propósito de separar seqüências de vogais.

d) Espaço em branco: Algumas vezes um espaço em branco não indica uma quebra na palavra. Por exemplo, caso seja decidido tratar database como sendo uma palavra, uma outra maneira de tratá-la como sendo uma palavra é escrevendo-a na forma data base. Destacam-se os números de telefone como sendo casos mais comuns dessa condição, onde um número como 9365 1873 é considerado uma única

palavra. Há também vários nomes constituídos de múltiplas partes, como New York e San Francisco, que são reconhecidos como uma única palavra. Especialmente torna-se difícil o caso em que este problema interage com hifenização, como demonstrado na expressão a seguir: "the New York-New Haven railroad". Neste caso, o hífen não expressa agrupamento com as "palavras gráficas" imediatamente adjacentes – o tratamento de York-New como unidade semântica seria um enorme equívoco.

Para as melhores características possíveis, deve-se sempre customizar o tokenizador para o texto a ser avaliado – caso contrário um trabalho extra pode ser requerido após os *token* terem sido obtidos. (WEIS, 2004)

Além da identificação de termos simples há ainda a identificação de termos compostos. Para WIVES, a identificação de termos compostos é o processo que determina as palavras que estão próximas dentro do texto e que as transforma em novos termos. Esses termos compostos são chamados de frases-termo.

3.6.2.3 Remoção dos vocábulos auxiliares

O processo de remoção dos vocábulos, conhecido como *stoplist* ou *stopword*, consiste em remover dos documentos termos que sejam pouco representativos para a sua descrição. É um dos primeiros passos do pré-processamento. São palavras auxiliares ou conectivas (e, para, a, eles) e que não fornecem nenhuma informação discriminativa na expressão do conteúdo dos textos (EBECKEN, LOPES e COSTA, 2003). Além disso, são palavras que ocorrem freqüentemente em todos os textos. Uma vez que elas muito comuns no conteúdo do documento, elas podem ser removidas do documento, para fins de categorização.

3.6.2.4 Stemming

O *stemming* (*lemmatization*) é o processo de unir as diferentes variações de uma palavra a sua representação comum, a raiz.

O *stemmer* pode ser de duas formas: o *Stemmer* Flexivo e o *Stemmer* para a Raiz. *Stemmer* flexivo é quando a normalização é confinada a regularizar variantes gramaticais como singular/plural e presente/passado. Na terminologia lingüística, isto é chamado "análise morfológica". (WEISS, 2004).

O *Stemming* para a Raiz tem como objetivo reduzir drasticamente o número de tipos de palavras dentro de uma coleção, fazendo assim uma distribuição estatística mais segura. (WEISS, 2004)

Existem diversos métodos de *stemming* desenvolvidos para a língua inglesa, tais como o Stemmer S, Porter e Lovins.

Para o algoritmo de classificação o uso do *stemming*, pode prover um pequeno benefício. (LEWIS, 2004).

3.6.2.5 Criação da *Bag of Words*

As características dos documentos são *tokens* ou termos que eles contêm. Se nenhum tipo de análise lingüística profunda é feita no conteúdo dos documentos, pode-se escolher descrever cada documento pelas características representadas pelos mais freqüentes *tokens*.

O conjunto coletivo de características é tipicamente chamado de dicionário. Os *tokens* ou palavras no dicionário formam a base para a criação da tabela de dados numéricos correspondentes a coleção de documentos. Cada linha é um documento, e cada coluna representa uma característica.

No modelo mais simples de dados simplesmente checa-se a presença ou ausência das palavras, e a entrada da célula será binária correspondente ao documento e a palavra. O dicionário de palavras cobre todas as possibilidades e corresponde ao número de colunas na tabela. As células terão um ou zero, dependendo se as palavras são encontradas no documento.

Como a meta é a predição, precisa-se de mais uma coluna para a resposta (ou classe) correta para cada documento. Preparando os dados para o método de aprendizagem, esta informação estará disponível dos rótulos de documento.

O rótulo é geralmente binário, e a menor classe é quase sempre a de interesse. Em vez de gerar um dicionário global para as duas classes, nos podemos considerar somente as palavras encontradas na classes que tentaremos predizer. Se esta classe é de longe menor que a classe negativa, o que é típico, o dicionário local será de longe menor que o dicionário global.

Na tabela gerada cada coluna representa um termo. As linhas representam os documentos. Um exemplo pode ser visto na Figura 7

Figura 7 - Ilustração de Bag Of Words

	Term₁	Term_k
d₁	a₁₁	a_{1k}
...
d_j	a_{j1}	a_{jk}

Fonte: LOPES, 2004

Para melhorar os resultados da predição, são realizadas transformações adicionais na tabela.

4 ESCOLHA DA METODOLOGIA DE CATEGORIZAÇÃO DE TEXTOS, SOFTWARES E ALGORITMOS A SEREM UTILIZADOS

4.1 ESCOLHA DA METODOLOGIA, ALGORITMOS E SOFTWARES

A metodologia empregada para a categorização dos textos foi a categorização automática de documentos textuais, usando a frequência estatística dos termos, pois além de ser uma abordagem mais trabalhada na literatura, ela apresenta bons resultados para a categorização. Os algoritmos utilizados são os categorizadores bayesiano e linear, implementados no sistema Aîuri.

4.2 ALGORITMOS UTILIZADOS

Os algoritmos utilizados pelo sistema Aîuri são o categorizadoresres bayesiano e linear.

4.3 Descrição da metodologia de aplicação do conjunto

Os passos da aplicação do conjunto metodologia, algoritmos e software selecionado para a aplicação da base de textos do Laboratório de Energia dos Ventos, são descritos no capítulo a seguir.

O software escolhido foi o sistema Aîuri, desenvolvido na COPPE/NACAD, pois além de ser um software acadêmico, ele oferece diversas vantagens, tais como a possibilidade de trabalhar localmente ou em Grid, como a possibilidade de sofrer expansões.

5 APLICAÇÃO AO CASO DO LABORATÓRIO DE ENERGIA DOS VENTOS

A base de dados textuais do Laboratório de Energia dos Ventos da Universidade Federal Fluminense (LEV/UFF), laboratório este que realiza estudos e projetos na área de energia, é um caso típico de armazenamento de arquivos digitais em grande quantidade, arquivos estes geralmente em formato PDF e disponíveis na Internet.

A facilidade atual de se obter documentos digitais, assim como as crescentes áreas de atuação e de interesse do Laboratório, gerou uma grande massa de arquivos textuais, em sua grande maioria nos idiomas inglês e português, ocupando algo em total de aproximadamente 22 GB de espaço em disco.

A recuperação de documentos importantes para um determinado assunto é, atualmente, extremamente ineficaz e ineficiente. Estas características fazem com que o caso da base textual do Laboratório seja um excelente caso para a verificação se a aplicação do conjunto selecionado de metodologia, algoritmo e software podem auxiliar no tratamento informacional necessário ao Laboratório.

5.1 COLETA DOS TEXTOS

Os documentos textuais que foram utilizados para a elaboração desta dissertação foram recolhidos do Laboratório de Energia dos Ventos (LEV/UFF). Como se tratava de uma base com uma grande quantidade de arquivos de diversos formatos, além de diferentes características (material multimídia, em diversos idiomas, diferentes tamanhos), recolhida durante anos de pesquisa, foi necessário um grande trabalho de refinamento e seleção dos dados que foram utilizados na dissertação.

A base utilizada é formada pelos documentos armazenados pelo diretor do Laboratório desde meados da década de 90 até os dias atuais, resultando em aproximadamente 12.000 arquivos em diversos formatos, que contenham informação relevante para o usuário.

Cabe ressaltar a forma como toda esta coleção foi gerada, textos totalmente desestruturados possuindo fotos, desenhos, textos escritos a mão. Os textos

registrados são livres, isto é, escritos em linguagem natural irrestrita, sem formatos ou padrões pré-estabelecidos e sem um vocabulário controlado. Os textos não sofreram nenhum tipo de correção. Erros porventura contidos neles foram mantidos.

Para a realização da dissertação foram recolhidos arquivos no formato PDF nas línguas portuguesa e inglesa. Foi realizada uma filtragem nos arquivos existentes, removendo arquivos duplicados, não pertencentes as categorias que se deseja trabalhar, corrompidos, que não contenham caracteres (sejam formados basicamente por figuras), que não pertençam aos idiomas que esta dissertação irá abordar (Inglês e Português), assim como arquivos de extensões que não sejam de documentos textuais.

Após a conclusão desta etapa, iniciou-se a colocação de nomes de arquivos condizentes ao seu conteúdo, visto que como foi uma base formada por arquivos “baixados” da Internet, não havia nenhuma padronização para os nomes de arquivo. Como padrão, colocou-se o título do documento, podendo ter ou não no título alguma observação pertinente. Superada esta etapa, dividiu-se a base em arquivos escritos na língua inglesa e os arquivos escritos em língua portuguesa. Esta etapa é importante pois faz-se um tratamento diferenciado dos arquivos por causa do idioma. Após a realização deste passo, converteu-se todos os arquivos para arquivos do tipo texto plano, necessário para o programa conseguir ler estes arquivos.

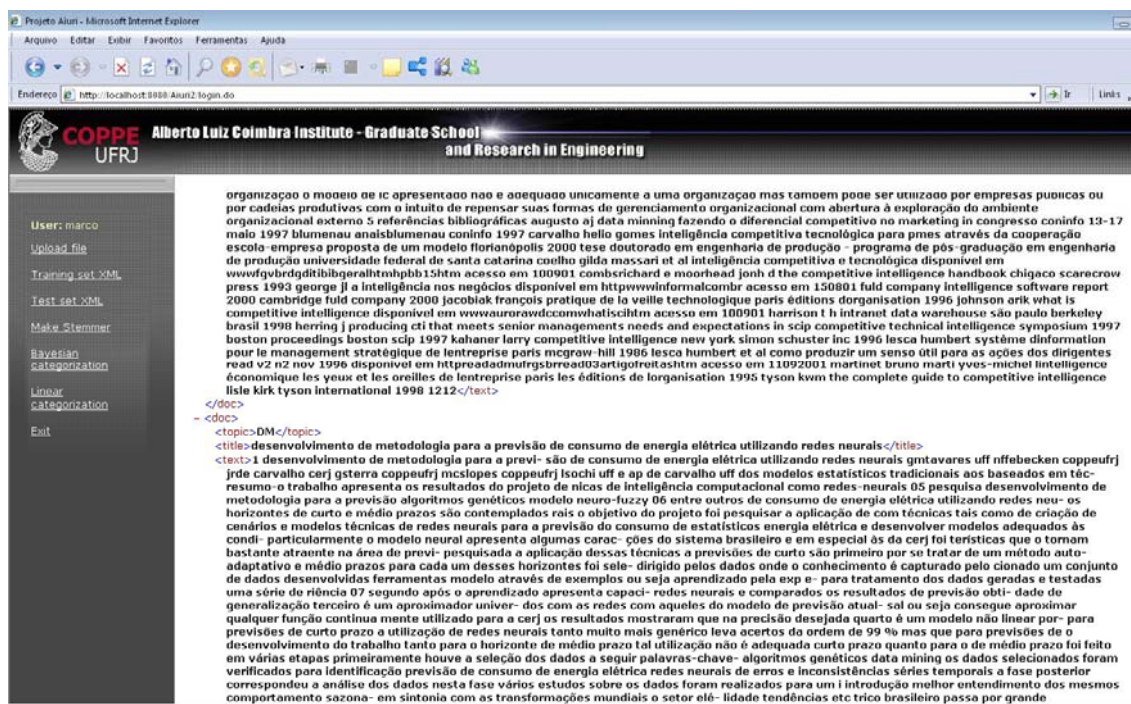
No total foram recolhidos 338 arquivos em português e 442 arquivos em inglês. Porém, nem todos os arquivos originalmente puderam ser utilizados, devido a problemas na conversão para texto (caracteres acentuados reproduzidos de forma errônea, textos contendo muitas palavras separadas, com muito “ruído” gerado na conversão) material que era somente iconográfico, arquivos corrompidos. Os documentos no Formato PDF variam de tamanho, tendo de 5 Kbytes até 68 Mbytes. Os arquivos, após convertidos para o formato texto, variaram de 2 kbytes até 756 kbytes para a base de arquivos em português e 1 Kbyte até 1.068 Kbytes para o idioma inglês. A listagem dos arquivos utilizados encontra-se nos apêndices 1 (Português) e 2 (Inglês).

Com os arquivos que foram selecionados da coleta anterior foi feita uma listagem (para cada base) com os nomes dos arquivos e uma breve descrição, visando a categorização destes arquivos por um especialista. As classes principais onde os arquivos foram alocados foram escolhidas pelo especialista. Nas tabelas abaixo estão descritos as classes e a quantidade de arquivo por cada classe.

Para que os arquivos pudessem ser manipulados pelo sistema, realizou, automaticamente, a conversão dos arquivos no formato texto (TXT) para o formato *Extensible Markup Language* (XML).

A Figura abaixo (8) mostra um exemplo de um arquivo no formato XML

Figura 8 – Exemplo de arquivo em formato XML



Na base contendo textos na língua portuguesa, as classes Mudança Climática (1 arquivo), Efeito Estufa (1 arquivo), Evasão (2 arquivos), Geração Distribuída (3 arquivos), Lablux (1 arquivo), Manutenção (2 arquivos), Operação (2 arquivos) não foram utilizadas por terem uma quantidade muito reduzida de arquivos. A tabela 1 mostra o quantitativo de arquivos em cada classe para o idioma Português.

Tabela 1 - Quantitativo de arquivos em cada classe para o idioma Português.

Português	
Área	Quantidade de arquivos
Data mining	6
Eficiência Energética	11
Energia	22
Energia Eólica	10
Equipamento Elétrico	14
Gerenciamento Energético	4

Pesquisa e Desenvolvimento	35
SIIF	68
Total:	170

Na base contendo textos na língua inglesa, as classes Automação (1 arquivo), Concessionária (2 arquivo), Dados (1 arquivo), Distribuição (2 arquivos), Engenharia (1 arquivo), Fotovoltaica (1 arquivo), Geração (3 arquivos), Gerenciamento (1 arquivo), GIS (2 arquivos), Hidrogênio (2 arquivos), Iluminação (1 arquivo), Infraestrutura (1 arquivo), Investimento (1 arquivo), Medição (2 arquivos), Meio ambiente (2 arquivos), Nanotecnologia (1 arquivo), Nuclear (2 arquivos), PCH (1 arquivo), Planejamento (1 arquivo), Política (2 arquivos), Política para as Renováveis (2 arquivos), Sensores (2 arquivos), Setor Elétrico (1 arquivo), Solar (1 arquivo), Turbina (2 arquivos) e UHE (1 arquivo) não foram utilizadas por terem uma quantidade muito reduzida de arquivos. A tabela 2 mostra o quantitativo de arquivos em cada classe para o idioma Inglês

Tabela 2 - Quantitativo de arquivos em cada classe para o idioma Inglês

Inglês	
Área	Quantidade de arquivos
Data mining	7
Educação	22
Eficiência Energética	43
Energia Distribuída	7
Energia	8
Energia Eólica	70
Geração Distribuída	18
Manutenção	5
Mercado	4
Mudanças Climáticas	8
Pesquisa e Desenvolvimento	13
Qualidade de Energia	8
Renováveis	18
Sistemas de Potência	4
Termelétrica	8
Tecnologia da Informação	5
Transmissão	5
Total:	253

5.2 PRÉ-PROCESSAMENTO

O pré-processamento é o conjunto de ações que visam preparar a base textual para um formato que o categorizador possa atuar sobre esta base.

5.2.1 CASE FOLDING

Para a padronização dos caracteres todos os textos foram convertidos para caixa baixa.

5.2.2 IDENTIFICAÇÃO DE TERMOS

Foi utilizada a função *java tokenizer*, da linguagem de programação utilizada na implementação do sistema

5.2.3 REMOÇÃO DOS VOCÁBULOS AUXILIARES

Neste trabalho foram utilizadas duas *stoplist*, uma para o idioma português, com 330 elementos e outra para o idioma inglês, com 366 elementos.

5.2.4 STEMMING

O sistema Aiuri gera o *stemming* somente para o idioma português. Sua apresentação é a seguinte: são duas colunas, onde a primeira coluna é o termo, e a segunda coluna é o termo após o processo de *stemming*.

O stemming utilizado pelo sistema foi o escrito por Orenge e Huyck (2001).

Na figura abaixo (9) tem-se um exemplo de criação do *stemming* pelo sistema.

Figura 9 – Stemming gerado pelo Sistema Aiuri

```
do do
projeto projet
foi foi
pesquisar pesquis
aplicação aplic
de de
con con

modelos model
técnicas tecnic
de de
redes red
neurais neur
para par
previsão previs
do do
consumo consua

energia energ
elétrica eletric
desenvolver desenvolv
modelos model
```

5.2.5 SELEÇÃO DAS CARACTERÍSTICAS

A principal característica, ou dificuldade, dos problemas de categorização é a alta dimensionalidade do espaço de características. O espaço de características nativo consiste nos termos únicos (palavras ou frases) que ocorrem nos documentos, que podem ter dezenas ou centenas de milhares de termos em uma coleção de textos de tamanho médio (YANG e PEDERSEN, 1997).

Os métodos de seleção de características incluem a remoção de termos não informativos de acordo com estatística do corpus, e a construção de novas características que combine características de baixo nível, (termos) dentro de dimensões ortogonais de alto nível (YANG e PEDERSEN, 1997).

A seleção das características (termos) melhora a capacidade de predição do categorizador. Alguns métodos de seleção de características são descritos abaixo.

Para o categorizador bayesiano utilizou-se a frequência binária, ou seja, se o termo existe ou não no documento. Para o categorizador linear utilizou-se a medida *tf-idf*.

5.2.5.1 Seleção por frequência relativa

A técnica da frequência relativa (também conhecida como *word count*) é uma das mais comuns no processo de seleção das características. A importância de um termo é dada pela quantidade de vezes que o termo aparece no texto.

5.2.5.2 Seleção por frequência do documento

A frequência do documento é o número de documentos que o termo ocorre. É computado, analisando a coleção inteira para determinar a frequência. A suposição básica é a de que termos raros não são informativos para a predição da categoria, ou não influenciam na performance global. A remoção destes termos diminui a dimensionalidade do espaço de características.

5.2.5.3 Produto da frequência do termo pela frequência inversa de documentos

Uma medida razoável da importância de um termo pode então ser obtida calculando, por exemplo, o produto da frequência do termo no documento pelo inverso da frequência de documentos em que o termo ocorre. A fórmula abaixo (A) mostra o peso tf-idf atribuído ao termo j é a frequência do termo (*word count*) modificada por uma escala da importância do termo. A escala é chamada de frequência inversa do documento, conforme vista na fórmula B. Ela simplesmente checa o número de documentos que contêm o termo j e inverte a escala.

$$tf - idf(j) = tf(j) * idf(j) \quad (A)$$

$$idf(j) = \log\left(\frac{N}{df_j}\right) \quad (B)$$

Quando o resultado do tf-idf é uma frequência alta significa que o termo aparece muitas vezes no documento, sendo considerado como não importante e, o tf-idf resultando em um valor baixo, próximo de 0, o termo é utilizado com pouca frequência e pode expressar uma importância maior no contexto do documento.

Bons termos descritores são os mais frequentes dentro de um texto porém infrequentes na coleção toda (frequência inversa) Os pesos devem ser normalizados para uma escala entre um e zero, para indicar a força relativa do termo descritor (Salton e McGill 1983 apud LOH, 2001).

5.2.6 VETORES ESPARSOS

Os documentos são convertidos em um formato de planilha onde cada linha corresponde ao documento, e cada coluna corresponde a uma palavra do dicionário. Este formato é conhecido como BOW (*Bag of Words*). As células individuais na planilha são preenchidas com a contagem da frequência (número de vezes que a palavra aparece no documento).

Tipicamente, o número de palavras (colunas) é muito grande e para muitos documentos (linhas), a maioria das palavras não se aplicam e o valor da célula é zero. Consequentemente é muito mais eficiente guardar somente a informação das células não-zero (o número da célula e o seu valor). Este é o vetor esparso.

O formato do arquivo do vetor esparso é o seguinte:

- Todos os pares não-zero para o documento aparece na mesma linha.
- Os pares não-zeros para cada documento devem estar em ordem crescente dos números da coluna
- Cada par não-zero é formatado como primeiro o número da coluna, um @ e depois a freqüência.

5.2.6.1 Vetores rotulados e não-rotulados

Para treinar o classificador, precisa-se etiquetar os documentos. Após obter o classificador, deve-se usá-lo para classificar documentos não-etiquetados. Os vetores são produzidos:

- Se os documentos são etiquetados, então as etiquetas aparecem na frente dos vetores correspondentes (separado com espaço em branco do par não-zero).
- Somente vetores binários são permitidos – documentos pertencem ou não a classe (etiqueta =1 ou a 0, respectivamente). O categorizador é construído para a classe positiva (casos com etiquetas =1)
- No arquivo vetor, todos os vetores deve ser do mesmo tipo (etiquetados ou não etiquetados)

5.2.6.2 Índice Invertido

O arquivo do vetor esparso mostra as características e as freqüências do vetor não-zero organizado da ordem do documento. O arquivo invertido contém exatamente a mesma informação, porém organizada na ordem das características. Para cada característica, ele lista os documentos no qual a característica é não-zero e também lista a freqüência relevante. Como esta informação é esparsa, o índice invertido usa várias ordens de ponteiros para armazenar esta informação, facilitando o acesso.

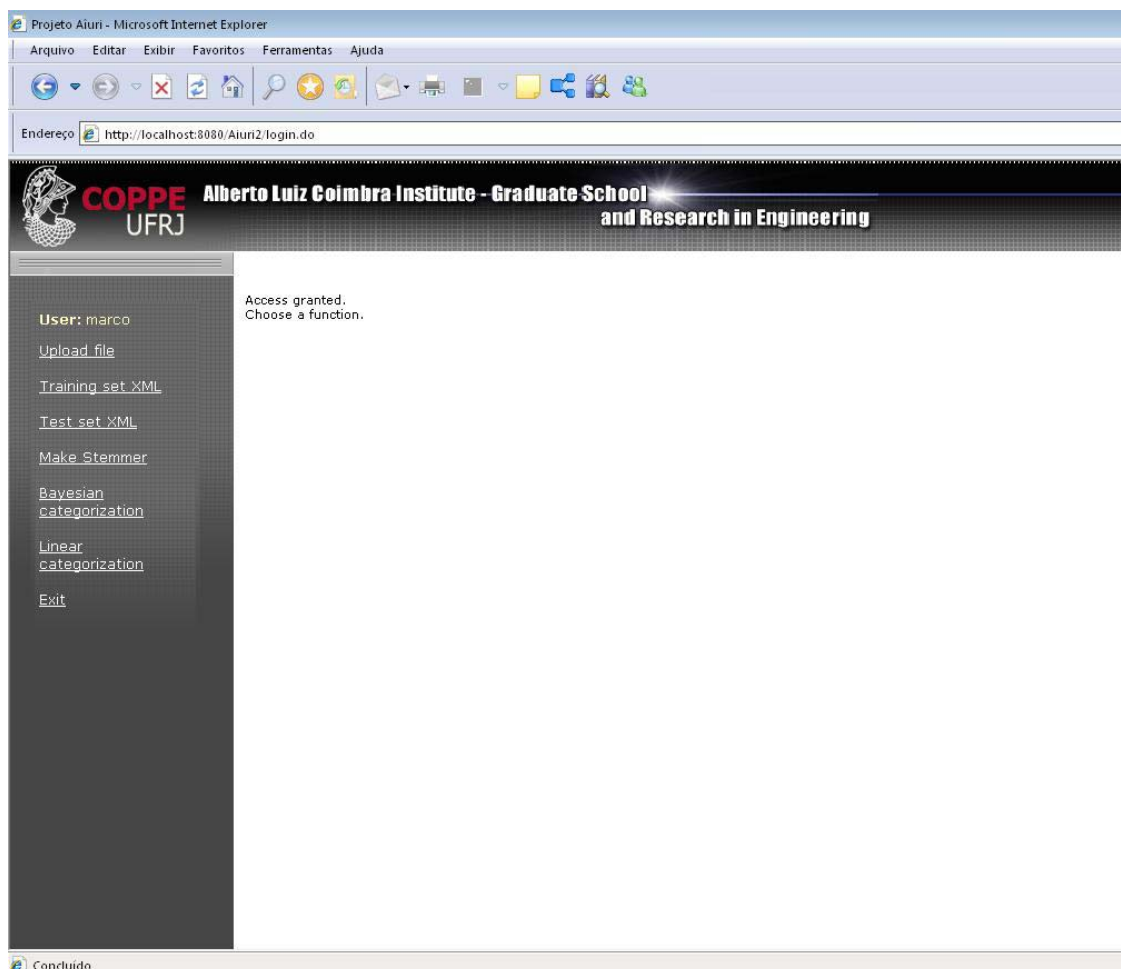
6 PROCESSAMENTO DOS DADOS

A etapa seguinte ao pré-processamento é a categorização dos documentos. Utiliza-se os resultados obtidos com o pré-processamento e aplica-se os algoritmos de classificação disponíveis no sistema Aîuri.

O sistema Aîuri é um sistema acadêmico, implementado em linguagem Java pelo COPPE/NACAD e possui as seguintes características: permite a utilização em dois ambientes: Computação em GRID e Local; interface *Web*; arquitetura servidor, conversão automática de documentos para XML; além de sua facilidade de expansão. A figura 10, abaixo, mostra a interface de entrada do sistema.

Devido as estas características, além da implementação de dois categorizadores (bayesiano e linear), o sistema Aîuri foi software escolhido para a categorização das bases do Laboratório.

Figura 10 - Tela de apresentação do Sistema Aiuri



A classificação foi realizada nos dois algoritmos disponíveis no sistema, o Bayesiano e o Linear. Para ambos foram utilizados o mesmo conjunto de treinamento e teste, buscando manter a proporção de 2/3 dos documentos de cada classe para treino e 1/3 para teste. Os arquivos de treino e de teste de cada classe foram selecionados de forma aleatória. As tabelas abaixo (3 e 4) mostra a proporção de arquivos utilizados para treinamento e teste dos algoritmos.

Tabela 3: Quantidade de arquivos utilizados em cada classe para treinamento e teste dos algoritmos – Idioma Português.

Português			
Área	Quantidade de arquivos	Treino	Teste
Data mining	6	4	2
Eficiência Energética	10	8	2
Energia	22	15	7
Energia Eólica	10	7	3
Equipamento Elétrico	14	10	4
Gerenciamento Energético	4	3	1
Pesquisa e Desenvolvimento	33	24	9
SIIF	68	45	23

Tabela 4: Quantidade de arquivos utilizados em cada classe para treinamento e teste dos algoritmos – Idioma Inglês.

Inglês			
Área	Quantidade de arquivos	Treino	Teste
Data mining	7	5	2
Educação	22	16	6
Eficiência Energética	43	30	13
Energia Distribuída	7	5	2
Energia	8	6	2
Energia Eólica	65	44	21
Geração Distribuída	18	12	6
Manutenção	5	4	1
Mercado	4	3	1
Mudanças Climáticas	8	6	2
Pesquisa e Desenvolvimento	13	9	4
Qualidade de Energia	8	6	2
Renováveis	18	13	5
Sistemas de Potência	4	3	1
Termelétrica	8	6	2
Tecnologia da Informação	5	4	1
Transmissão	5	4	1

Após a conversão dos arquivos de treino e teste, parte-se para o processo de criação do *stemming*. O algoritmo de *stemming* para a língua portuguesa utilizado no portal foi o Removedor de Sufixos da Língua Portuguesa, uma adaptação do *stemming* de Porter para o Português do Brasil.

Após o Processo de criação do *stemming*, ajusta-se os parâmetros para a categorização, como o tamanho do dicionário a classe que se deseja categorizar, o carregamento da lista de *stopwords* e a lista com o *stemming*. Outros ajustes podem ser feitos para melhorar a performance do categorizador em uso.

6.1 CLASSIFICADOR BAYESIANO

O método mais óbvio para classificação é procurar diretamente a probabilidade das palavras no documento.

O sistema Aîuri possui um classificador binário bayesiano. Aplica-se o classificador aos novos vetores e divide os documentos correspondentes aos vetores em duas partes: os classificados como positivo e os classificados como negativo. O sistema gera as duas listagens. A definição do classificador bayesiano encontra-se no item 3.5.8.

6.2 CLASSIFICADOR LINEAR

O sistema Aîuri possui um classificador Linear. Aplica-se o classificador linear para os novos vetores e divide os documentos correspondentes aos vetores em duas partes: os classificados como positivo e os classificados como negativo. O sistema gera as duas listagens. A definição do classificador linear encontra-se no item 3.5.9.

7 AVALIAÇÃO DOS RESULTADOS

O pós-processamento dos dados consiste da fase das descobertas efetuadas pela etapa de processamento dos dados e da visualização dos resultados encontrados.

Métricas de avaliação de resultados, ferramentas de visualização e conhecimento de especialistas ajudam a consolidar os resultados.

Para a avaliação da eficácia da categorização atribuída pelos classificadores aos documentos, usam-se comumente as medidas Revocação, Precisão, Média F e a Matriz de Confusão.

7.1 PRECISÃO

É a relação das atribuições corretas pelo sistema dividido pelo número total de atribuições do sistema. É obtido pela fórmula:

$$\text{Precisão} = \frac{\text{Número de itens relevantes recuperados}}{\text{Número total de itens recuperados}}$$

7.2 REVOCAÇÃO

É definido como sendo a relação das corretas atribuições de classes pelo sistema dividido pelo número total de atribuições corretas. É obtido pela fórmula:

$$\text{Revocação} = \frac{\text{Número de itens relevantes recuperados}}{\text{Número de itens relevantes na coleção}}$$

7.3 MÉDIA F

É definido como a média harmônica de precisão e revocação. É frequentemente usado para medir o desempenho de um sistema quando um único número é preferido.

$$\text{Média F} = \frac{2}{1/\text{precision} + 1/\text{recall}}$$

7.4 MATRIZ DE CONFUSÃO

A matriz de confusão mostra como os erros de classificação foram distribuídos. O modelo é considerado bom quando os elementos da diagonal principal da matriz são altos, enquanto os outros são próximos ou, iguais a zero. A figura 11 exemplifica uma matriz de confusão. Nesta dissertação a classe representada na figura 11 como classe A é a classe que está sendo analisada e a classe B é o somatório das outras classes.

Figura 11 – Exemplo de uma matriz de confusão

		Classe prevista	
		A	B
Classe real	A	Quantidade de registros da classe A classificados como A	Quantidade de registros da classe A classificados como B
	B	Quantidade de registros da classe B classificados como A	Quantidade de registros da classe B classificados como B

Fonte: Pereira e Santos (2004)

7.5 RESULTADOS PARA A BASE EM IDIOMA PORTUGUÊS

7.5.1 Classe Data Mining

Tabela 5 – Resultados para a classe Data Mining no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	18.181	100	30.76
50	Linear	50	50	50
100	Bayesiano	12.5	50	20
100	Linear	100	50	66.66
200	Bayesiano	50	50	50
200	Linear	100	50	66.66
500	Bayesiano	100	50	66.66
500	Linear	100	0	0
1000	Bayesiano	100	50	66.66
1000	Linear	100	0	0

Para os documentos pertencentes a classe *data mining* em português os classificadores naïve bayes e linear apresentaram resultados semelhantes, sendo que o linear foi um pouco melhor pois obteve um melhor resultado em várias opções de tamanho de dicionário. Como possuía poucos arquivos de treino e teste (4 e 2, respectivamente) um erro de categorização de apenas 1 arquivo produz resultados ruins. Nos erros de precisão inseriu arquivos das classes de Pesquisa e Desenvolvimento (principalmente), Energia e Energia Eólica.

A matriz de confusão para o resultado em negrito na tabela 5 foi a seguinte:

Classe Data Mining – Resultado do Classificador Linear com dicionário de 100 termos	
1	1
0	49

7.5.2 Classe Eficiência Energética

Tabela 6 – Resultados para a classe Eficiência Energética no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	11.11	50	18.18
50	Linear	33.33	50	40
100	Bayesiano	9.09	50	15.38
100	Linear	20	50	28.57

200	Bayesiano	11.11	50	18.18
200	Linear	0	0	0
500	Bayesiano	9.09	50	15.38
500	Linear	100	0	0
1000	Bayesiano	14	50	22
1000	Linear	100	0	0

Para os documentos pertencentes a classe eficiência energética em português os classificador linear obteve resultados um pouco melhores, principalmente com o uso de dicionários pequenos. Nos erros de precisão inseriu arquivos de todas as classes, principalmente energia.

A matriz de confusão para o resultado em negrito na tabela 6 foi a seguinte:

Classe Eficiência Energética – Resultado do Classificador Linear com dicionário de 50 termos	
1	1
2	47

7.5.3 Classe Energia

Tabela 7 – Resultados para a classe Energia no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	54.54	85.71	66.66
50	Linear	50	28.57	36.36
100	Bayesiano	41.66	71.48	52.63
100	Linear	57.14	57.14	57.14
200	Bayesiano	11.11	50	18.18
200	Linear	66.66	57.14	61.53
500	Bayesiano	57.14	57.14	57.14
500	Linear	80	57.14	66.66
1000	Bayesiano	66.66	57.142	61.53
1000	Linear	100	14.28	25

Para os documentos pertencentes a classe energia em português os classificador linear obteve resultados um pouco melhores, sendo que o seu melhor resultado foi com o uso de um dicionário de 500 palavras. O classificador Bayesiano conseguiu um bom resultado com o uso de um dicionário pequeno. Nos erros de precisão inseriu arquivos das classes *data mining*, eficiência energética e energia eólica.

A matriz de confusão para o resultado em negrito na tabela 7 foi a seguinte:

Classe Energia – Resultado do Classificador Bayesiano com dicionário de 50 termos	
6	1
5	39

7.5.4 Classe SIIF

Tabela 8 – Resultados para a classe SIIF no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	100	95.45	97.67
50	Linear	95.45	95.45	95.45
100	Bayesiano	100	72.72	84.21
100	Linear	95.45	95.45	95.45
200	Bayesiano	100	72.72	84.21
200	Linear	100	95.45	97.67
500	Bayesiano	100	72.72	84.21
500	Linear	100	95.45	97.67
1000	Bayesiano	100	72.72	84.21
1000	Linear	100	95.45	97.64

A classe SIIF foi onde os categorizadores obtiveram um ótimo desempenho. Deve-se o fato a qualidade dos dados, pois eram relatórios bem padronizados. Quando classificou erroneamente colocou um arquivo da classe energia eólica.

A matriz de confusão para o resultado em negrito na tabela 8 foi a seguinte:

Classe SIIF – Resultado do Classificador Bayesiano com	
--	--

dicionário de 50 termos	
21	2
0	28

7.5.5 Classe Equipamento Elétrico

Tabela 9 – Resultados para a classe Equipamento Elétrico no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	50	25	33.33
50	Linear	100	25	40
100	Bayesiano	100	25	40
100	Linear	100	25	40
200	Bayesiano	100	25	40
200	Linear	100	25	40
500	Bayesiano	100	25	40
500	Linear	100	50	66.66
1000	Bayesiano	100	25	40
1000	Linear	100	75	85.71

Nesta classe o melhor resultado foi obtido pelo categorizador linear utilizando um dicionário grande. Errou na precisão por classificar um documento pertencente a classe energia nele. Em seu melhor resultado alocou corretamente 3 arquivos.

A matriz de confusão para o resultado em negrito na tabela 9 foi a seguinte:

Classe Equipamento Elétrico – Resultado do Classificador Linear com dicionário de 1000 termos	
3	1
0	47

7.5.6 Classe Gerenciamento Energético

Tabela 10 – Resultados para a classe Gerenciamento Energético no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	0	0	0
50	Linear	50	100	66
100	Bayesiano	100	0	0
100	Linear	100	0	0
200	Bayesiano	100	0	0
200	Linear	100	0	0
500	Bayesiano	100	0	0
500	Linear	100	0	0
1000	Bayesiano	100	0	0
1000	Linear	100	0	0

O único resultado aproveitável foi obtido pelo categorizador linear utilizando um dicionário de tamanho pequeno. Possui poucos registros, o que inviabilizou melhores resultados. No melhor resultado colocou um arquivo da classe pesquisa e desenvolvimento como sendo desta classe.

A matriz de confusão para o resultado em negrito na tabela 10 foi a seguinte:

Classe Energético – Gerenciamento	
Resultado do Classificador Linear com dicionário de 50 termos	
1	0
1	49

7.5.7 Classe Pesquisa e Desenvolvimento

Tabela 11 – Resultados para a classe Pesquisa e Desenvolvimento no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	50	88.88	64
50	Linear	54.54	66.66	60
100	Bayesiano	50	88.88	64
100	Linear	100	88.88	94.11
200	Bayesiano	50	88.88	64

200	Linear	75	100	85.71
500	Bayesiano	47.05	88.88	88.88
500	Linear	88.88	88.88	88.88
1000	Bayesiano	53.33	88.88	66.66
1000	Linear	100	88.88	94.11

O classificador linear, utilizando um dicionário pequeno e um grande, foi o que obteve os melhores resultados nesta classe. Quando erro, alocou arquivos das classes gerenciamento e energia.

A matriz de confusão para o resultado em negrito na tabela 11 foi a seguinte:

Classe Pesquisa e Desenvolvimento – Resultado do Classificador Linear com dicionário de 100 termos	
8	1
0	42

7.5.8 Classe Energia Eólica

Tabela 12 – Resultados para a classe Energia Eólica no idioma Português

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	50	66.66	57.14
50	Linear	50	66.66	57.14
100	Bayesiano	50	66.66	57.14
100	Linear	40	66.66	50
200	Bayesiano	50	66.66	57.14
200	Linear	100	0	0
500	Bayesiano	50	66.66	57.14
500	Linear	0	0	0
1000	Bayesiano	66.66	66.66	66.66
1000	Linear	50	33.33	40

O classificador Bayesiano, utilizando um dicionário grande, foi o que obteve os melhores resultados nesta classe. Quando a configuração com um dicionário de 100 palavras errou, alocou arquivo da classe energia.

A matriz de confusão para o resultado em negrito na tabela 12 foi a seguinte:

Classe Energia Eólica – Resultado do Classificador Bayesiano com dicionário de 1000 termos	
2	1
1	47

7.6 RESULTADOS PARA A BASE EM IDIOMA INGLÊS

7.6.1 Classe Data Mining

Tabela 13 – Resultados para a classe Data Mining no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	14	50	22.22
50	Linear	0	0	0
100	Bayesiano	0	0	0
100	Linear	0	0	0
200	Bayesiano	0	0	0
200	Linear	0	0	0
500	Bayesiano	0	0	0
500	Linear	100	0	0
1000	Bayesiano	0	0	0
1000	Linear	100	0	0

Os categorizadores não obtiveram êxito na classificação dos arquivos da classe *data mining* em inglês. No único resultado não zero que obtiveram, o classificador bayesiano alocou nesta classe arquivos das classes eficiência energética, energia, manutenção, mercado, pesquisa e desenvolvimento e qualidade de energia. Deve-se a esta classificação ruim o fato dos arquivos desta classe estarem diretamente ligados a outras, são aplicações de computação na área de energia.

A matriz de confusão para o resultado em negrito na tabela 13 foi a seguinte:

Classe Data Mining – Resultado do Classificador Bayesiano com dicionário de 50 termos	
1	1
6	64

7.6.2 Classe Eficiência Energética

Tabela 14 – Resultados para a classe Eficiência Energética no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	33.33	46.15	38.70
50	Linear	60	46.15	52.17
100	Bayesiano	25	23.07	24
100	Linear	66.66	61.53	64
200	Bayesiano	21.05	30.76	25.0
200	Linear	81.81	69.23	75
500	Bayesiano	20	30.76	24.24
500	Linear	75	46.15	57.14
1000	Bayesiano	20	30.76	24.24
1000	Linear	83.33	38.46	52.63

O melhor resultado da classe eficiência energética para o idioma inglês foi o obtido pelo categorizador linear utilizando um dicionário de 200 palavras. Geralmente foram alocados nesta classe arquivos pertencentes as classes energia distribuída, energia eólica, geração distribuída, manutenção, pesquisa e desenvolvimento, renováveis e termelétrica.

A matriz de confusão para o resultado em negrito na tabela 14 foi a seguinte:

Classe Eficiência Energética – Resultado do Classificador Linear com dicionário de 200 termos	
9	4
2	57

7.6.3 Classe Energia

Tabela 15 – Resultados para a classe Energia no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	0	0	0
50	Linear	25	50	33.33
100	Bayesiano	7.69	50	13.33
100	Linear	0	0	0
200	Bayesiano	10	50	16.66
200	Linear	100	0	0
500	Bayesiano	16.66	50	25
500	Linear	100	0	0
1000	Bayesiano	33.33	50	40
1000	Linear	100	0	0

Nesta classe os categorizadores não obtiveram êxito. O melhor resultado foi o obtido pelo classificador bayesiano, utilizando um grande dicionário. Ele alocou nesta classe arquivos pertencentes em sua maioria a classe energia eólica.

A matriz de confusão para o resultado em negrito na tabela 15 foi a seguinte:

Classe Energia – Resultado do Classificador Bayesiano com dicionário de 1000 termos	
1	1
2	68

7.6.4 Classe Educação

Tabela 16 – Resultados para a classe Educação no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	50	83.33	62.5
50	Linear	100	83.33	90.90
100	Bayesiano	50	83.33	62.5
100	Linear	83.33	83.33	83.33
200	Bayesiano	45.45	83.33	58.82

200	Linear	100	83.33	90.90
500	Bayesiano	50	83.33	62.5
500	Linear	100	100	100
1000	Bayesiano	71.42	83.33	76.92
1000	Linear	100	66.66	80

O melhor resultado foi o obtido pelo classificador linear utilizando um dicionário de 500 palavras. Nos erros cometidos alocou arquivos da classe renováveis (principalmente) assim como energia eólica, manutenção, eficiência energética e pesquisa e desenvolvimento.

A matriz de confusão para o resultado em negrito na tabela 16 foi a seguinte:

Classe Educação – Resultado do Classificador Linear com dicionário de 500 termos	
6	0
0	66

7.6.5 Classe Energia Distribuída

Tabela 17 – Resultados para a classe Energia Distribuída no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	33.33	100	50
50	Linear	50	100	66.66
100	Bayesiano	11.11	50	18.18
100	Linear	50	100	66.66
200	Bayesiano	20	100	33.33
200	Linear	25	50	33.33
500	Bayesiano	12.50	50	20
500	Linear	0	0	0
1000	Bayesiano	14.28	50	22.22
1000	Linear	0	0	0

Os melhores resultados obtidos nesta classe foram com o classificador linear utilizando dicionários pequenos. Geralmente alocou arquivos das classes geração

distribuída (frequentemente) eficiência energética, energia eólica, manutenção, pesquisa e desenvolvimento.

A matriz de confusão para o resultado em negrito na tabela 17 foi a seguinte:

Classe Energia Distribuída – Resultado do Classificador Linear com dicionário de 50 termos	
2	0
2	68

7.6.6 Classe Manutenção

Tabela 18 – Resultados para a classe Manutenção no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	25	100	40
50	Linear	0	0	0
100	Bayesiano	25	100	40
100	Linear	0	0	0
200	Bayesiano	33.33	100	50
200	Linear	100	0	0
500	Bayesiano	0	0	0
500	Linear	100	0	0
1000	Bayesiano	0	0	0
1000	Linear	100	0	0

Os categorizadores não obtiveram êxito nesta classe. Alocaram arquivos pertencentes as classes eficiência energética, energia eólica e qualidade de energia.

A matriz de confusão para o resultado em negrito na tabela 18 foi a seguinte:

Classe Manutenção – Resultado do Classificador Bayesiano com dicionário de 200 termos	
1	0
2	69

7.6.7 Classe Pesquisa e Desenvolvimento

Tabela 19 – Resultados para a classe Pesquisa e Desenvolvimento no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	6.66	25	10.52
50	Linear	0	0	0
100	Bayesiano	8.33	25	12.5
100	Linear	0	0	0
200	Bayesiano	12.5	25	16.66
200	Linear	0	0	0
500	Bayesiano	16.66	25	20
500	Linear	0	0	0
1000	Bayesiano	14.28	25	18.18
1000	Linear	0	0	0

Os classificadores não obtiveram êxito nesta categorização. Alocaram arquivos das classes eficiência energética, energia distribuída, energia eólica (principalmente), geração distribuída, manutenção, qualidade de energia, renováveis e transmissão.

A matriz de confusão para o resultado em negrito na tabela 19 foi a seguinte:

Classe Pesquisa e Desenvolvimento – Resultado do Classificador Bayesiano com dicionário de 500 termos	
1	3
5	63

7.6.8 Classe Energia Eólica

Tabela 20 – Resultados para a classe Energia Eólica no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	80	76.19	78.04
50	Linear	76	90	82.60
100	Bayesiano	73.91	80.95	77.27
100	Linear	72	85.71	78.26

200	Bayesiano	55.17	76.19	64
200	Linear	81.81	85.71	83.72
500	Bayesiano	43.24	76.19	55.17
500	Linear	85	80.95	82.92
1000	Bayesiano	41.02	76.19	53.33
1000	Linear	85	80.95	82.92

O classificador linear utilizando um dicionário de 200 palavras obteve o melhor resultado. Quando errou, alocou arquivos das classes geração distribuída, pesquisa e desenvolvimento, renováveis e termelétrica.

A matriz de confusão para o resultado em negrito na tabela 20 foi a seguinte:

Classe Energia Eólica – Resultado do Classificador Linear com dicionário de 200 termos	
18	3
4	47

7.6.9 Classe Geração Distribuída

Tabela 21 – Resultados para a classe Geração Distribuída no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	40	75	52.17
50	Linear	41.66	62	50
100	Bayesiano	35.71	62.5	45.45
100	Linear	55.55	62.5	58.82
200	Bayesiano	38.46	62.5	47.61
200	Linear	55.55	62.5	58.82
500	Bayesiano	55.55	62.5	58.82
500	Linear	50	62.5	55.55
1000	Bayesiano	71.42	62.5	66.66
1000	Linear	62.5	62.5	62.5

O classificador Bayesiano utilizando um dicionário de 1000 palavras obteve o melhor resultado. Quando errou, alocou arquivos das classes energia distribuída, pesquisa e desenvolvimento e renováveis.

A matriz de confusão para o resultado em negrito na tabela 21 foi a seguinte:

Classe Geração Distribuída – Resultado do Classificador Bayesiano com dicionário de 1000 termos	
3	3
4	62

7.6.10 Classe Mercado

Tabela 22 – Resultados para a classe Mercado no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	0	0	0
50	Linear	100	0	0
100	Bayesiano	0	0	0
100	Linear	100	0	0
200	Bayesiano	0	0	0
200	Linear	100	0	0
500	Bayesiano	0	0	0
500	Linear	100	0	0
1000	Bayesiano	100	0	0
1000	Linear	100	0	0

Nenhum dos categorizadores conseguiu êxito nesta classe. Quando alocou arquivos, alocou arquivos das classes eficiência energética, pesquisa e desenvolvimento (principalmente), geração distribuída e renováveis.

A matriz de confusão para o resultado em negrito na tabela 22 foi a seguinte:

Classe Mercado – Resultado do Classificador Bayesiano com dicionário de 50 termos	
0	1

4	67
---	----

7.6.11 Classe Mudanças Climáticas

Tabela 23 – Resultados para a classe Mudanças Climáticas no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	18.18	100	30.76
50	Linear	50	50	50
100	Bayesiano	8.33	50	14.28
100	Linear	0	0	0
200	Bayesiano	8.33	50	14.28
200	Linear	0	0	0
500	Bayesiano	0	0	0
500	Linear	0	0	0
1000	Bayesiano	0	0	0
1000	Linear	0	0	0

Nenhum dos categorizadores conseguiu muito êxito nesta classe. O melhor resultado foi obtido pelo classificador linear, utilizando um dicionário pequeno. Quando o classificador linear utilizando esta configuração errou, alocou arquivo da classe energia eólica.

A matriz de confusão para o resultado em negrito na tabela 23 foi a seguinte:

Classe Mudanças Climáticas – Resultado do Classificador Linear com dicionário de 50 termos	
1	1
1	69

7.6.12 Classe Qualidade de Energia

Tabela 24 – Resultados para a classe Qualidade de Energia no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	28.57	100	44.44
50	Linear	0	0	0
100	Bayesiano	25	50	33.33
100	Linear	0	0	0
200	Bayesiano	20	50	28.57
200	Linear	100	0	0
500	Bayesiano	20	50	28.57
500	Linear	100	0	0
1000	Bayesiano	33.33	50	40
1000	Linear	100	0	0

Nenhum dos categorizadores conseguiu muito êxito nesta classe. O melhor resultado foi obtido pelo classificador Bayesiano, utilizando um dicionário pequeno. Quando o classificador Bayesiano utilizando esta configuração errou, alocou arquivo das classes eficiência energética, manutenção, pesquisa e desenvolvimento, termelétrica e transmissão.

A matriz de confusão para o resultado em negrito na tabela 24 foi a seguinte:

Classe Qualidade de Energia – Resultado do Classificador Bayesiano com dicionário de 50 termos	
2	0
5	65

7.6.13 Classe Renováveis

Tabela 25 – Resultados para a classe Renováveis no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	15.78	60	25
50	Linear	60	60	60
100	Bayesiano	16.66	60	26.08
100	Linear	42.85	60	50
200	Bayesiano	10.52	40	16.66

200	Linear	25	20	22.22
500	Bayesiano	10.52	40	16.66
500	Linear	33.33	20	25
1000	Bayesiano	12.5	40	19.04
1000	Linear	100	0	0

Nenhum dos categorizadores conseguiu muito êxito nesta classe. O melhor resultado foi obtido pelo classificador linear, utilizando um dicionário de 50 palavras. Quando o classificador linear utilizando esta configuração errou, alocou arquivo das classes energia distribuída e transmissão.

A matriz de confusão para o resultado em negrito na tabela 25 foi a seguinte:

Classe Renováveis – Resultado do Classificador Linear com dicionário de 50 termos	
3	2
2	65

7.6.14 Classe Sistemas de Potência

Tabela 26 – Resultados para a classe Sistemas de Potência no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	0	0	0
50	Linear	0	0	0
100	Bayesiano	0	0	0
100	Linear	0	0	0
200	Bayesiano	0	0	0
200	Linear	100	0	0
500	Bayesiano	0	0	0
500	Linear	100	0	0
1000	Bayesiano	100	0	0
1000	Linear	100	0	0

Nenhum dos categorizadores conseguiu êxito nesta classe. Quando alocou arquivos, eles foram das classes eficiência energética, pesquisa e desenvolvimento, energia eólica, geração distribuída, manutenção, energia distribuída, e renováveis.

A matriz de confusão para o resultado em negrito na tabela 26 foi a seguinte:

Classe Sistemas de Potência – Resultado do Classificador Bayesiano com dicionário de 50 termos	
0	1
1	69

7.6.15 Classe Termelétrica

Tabela 27 – Resultados para a classe Termelétrica no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	50	100	66.66
50	Linear	100	50	66.66
100	Bayesiano	50	100	66.66
100	Linear	100	50	66.66
200	Bayesiano	66.66	100	80
200	Linear	100	50	66.66
500	Bayesiano	50	50	50
500	Linear	100	50	66.66
1000	Bayesiano	50	50	50
1000	Linear	100	50	66.66

O melhor resultado obtido para esta classe foi pelo categorizador Bayesiano utilizando um dicionário de 200 palavras. Quando o classificador Bayesiano utilizando esta configuração errou, alocou arquivo da classe eficiência energética.

A matriz de confusão para o resultado em negrito na tabela 27 foi a seguinte:

Classe Termelétrica – Resultado do Classificador Bayesiano com dicionário de 200 termos	
2	0

1	69
---	----

7.6.16 Classe Tecnologia da Informação

Tabela 28 – Resultados para a classe Tecnologia da Informação no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	0	0	0
50	Linear	100	100	100
100	Bayesiano	0	0	0
100	Linear	100	0	0
200	Bayesiano	0	0	0
200	Linear	100	0	0
500	Bayesiano	0	0	0
500	Linear	100	0	0
1000	Bayesiano	0	0	0
1000	Linear	100	0	0

O melhor resultado obtido para esta classe foi pelo categorizador linear utilizando um dicionário de 50 palavras. Alocou corretamente o arquivo pertencente a classe Tecnologia da Informação.

A matriz de confusão para o resultado em negrito na tabela 28 foi a seguinte:

Classe Tecnologia da Informação – Resultado do Classificador Linear com dicionário de 50 termos	
1	0
0	71

7.6.17 Classe Transmissão

Tabela 29 – Resultados para a classe Transmissão no idioma Inglês

Dicionário	Classificador	Precisão	Revocação	Média F
50	Bayesiano	16.66	100	28.57

50	Linear	100	100	100
100	Bayesiano	20	100	33.33
100	Linear	100	100	100
200	Bayesiano	16.66	100	28.57
200	Linear	100	100	100
500	Bayesiano	20	100	33.33
500	Linear	100	0	0
1000	Bayesiano	0	0	0
1000	Linear	100	0	0

O melhor resultado obtido para esta classe foi pelo categorizador linear utilizando diversas configurações de dicionário. Alocou corretamente o arquivo pertencente a classe Transmissão. Quando se aumentou o número de palavras do dicionário, seu desempenho decaiu.

A matriz de confusão para o resultado em negrito na tabela 29 foi a seguinte:

Classe Transmissão – Resultado do Classificador Linear com dicionário de 50 termos	
1	0
0	71

8 CONCLUSÕES

Nesta dissertação foi efetuada a categorização de arquivos pertencentes a um laboratório de pesquisa, arquivos estes nos idiomas inglês e português.

A possibilidade de permitir ao usuário que ele informe as classes que deseja em que seus documentos sejam alocados, ao invés de classes já informadas no sistema, permite ao usuário uma melhor organização.

Essa possibilidade do usuário em atribuir categorias permite a criação de categorias “idênticas”, pois energia distribuída e geração distribuída são praticamente sinônimos.

O uso da técnica de categorização automática de textos é um grande aliado na organização dos arquivos textuais. Neste trabalho mostrou ter bons resultados com o uso da técnica.

No *corpus* no idioma português percebe-se que classes como SIIF e equipamento elétrico que não apresentam uma correlação alta com as outras. Porém, arquivos de classes mais gerais, como Energia, foram classificados dentro de outras classes como Eficiência Energética, Energia Eólica, Pesquisa e Desenvolvimento e Data Mining. Percebeu-se a correlação entre os arquivos das classes Energia e Eficiência Energética; Gerenciamento e Pesquisa e Desenvolvimento; Energia e Data Mining.

No *corpus* no idioma Inglês, que possui mais arquivos e mais classes, percebe-se que classes como: Tecnologia da Informação, Educação, Transmissão, Mercado e Data Mining não apresentam uma correlação alta com as outras. Percebeu-se a correlação entre os arquivos das classes: Eficiência Energética e Manutenção; Eficiência Energética e Energia Distribuída; Eficiência Energética e Pesquisa e Desenvolvimento; Eficiência Energética e Termelétrica; Energia Distribuída e Geração Distribuída; Energia Distribuída e Pesquisa e Desenvolvimento; Manutenção e Qualidade de Energia; Pesquisa e Desenvolvimento e Qualidade; Geração Distribuída e Pesquisa e Desenvolvimento.

As classes eram desbalanceadas. Percebe-se que com uma base de treino e teste reduzidas para algumas classes os categorizadores não obtiveram êxito. Para uma melhor categorização duas coisas se mostraram fundamentais: que se tenha arquivos de treino e de teste em boa quantidade, além dos arquivos pertencentes a

uma classe terem características em comum entre eles, e que estas características sejam diferentes para os outros arquivos das outras classes.

Os testes indicaram que o melhor resultado foi geralmente obtido pelo classificador linear.

8.1 SUGESTÕES PARA TRABALHOS POSTERIORES

Devido a característica do sistema Aîuri permitir que novos categorizadores sejam adicionados, seria interessante a inserção de outros categorizadores, preferencialmente multiclassés.

Como o sistema é bem flexível, o acréscimo de outras ferramentas de mineração de textos seria bastante útil.

Além disto, o sistema poderá sofrer alterações para algumas etapas importantes, tais como o acréscimo de um processo que identifique “ruídos” em arquivos, a implementação da conversão de arquivos no formato PDF para o formato texto pelo sistema, a implementação de um processo de identificação de idioma do texto para a divisão da base por idiomas.

A realização de mais trabalhos que não sejam em bases de *benchmark* também é interessante.

REFERÊNCIAS BIBLIOGRÁFICAS

AL-SHALABI, RIYAD; KANAAN. GHASSAN; GHARAIBEH, MANAF H. **Arabic Text Categorization Using kNN Algorithm.** 2004. Disponível em: < www.ijicis.net/Vol6_No1%20No_1.pdf >. Acesso em: 20 Out 2006.

ALVES, ABEL. **IBM: gerenciando a confusão: empresa prevê que a avalanche de informação vai se tornar (quase) insuportável.** O GLOBO, Info etc., 23 out. 2006, pág. 4.

BAEZA-YATES, R., RIBEIRO NETO, B. **Modern information retrieval.** England : ACM press, 1999.

CRESTANI, F., PASI, G. **Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks.** Disponível em: < <http://www.cis.strath.ac.uk/~fabioc/papers/99-fuzzy-book.pdf> >. Acesso em: 09 ago 2005.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. A. Mineração de textos. In: . **Sistemas inteligentes: fundamentos e aplicação** [REZENDE, S. O. (Org.)]. Barueri Manole, 2003. cap. 13, p. 337-370.

FERNEDA, E. Recuperação da informação: análise sobre a contribuição da ciência da computação para a ciência da informação. São Paulo: USP, 2003. 147p. Tese (Ciências da Comunicação) – Escola de Comunicação e Arte da Universidade de São Paulo. Disponível em: < <http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf> >. Acesso em: 11 set. 2006.

GALHO, THAÍS SILVA. **Categorização automática de documentos de texto utilizando lógica difusa.** Gravataí: Universidade Luterana do Brasil, Curso de Ciência da Computação, 2003, 79p.

GONÇALVES, Lea Silvia Martins; REZENDE, Solange Oliveira. Categorização em Text mining. Disponível em <http://www.di.ubi.pt/~api/text_categorization.pdf>. Acesso em: 02 set. 2006.

INDURKHYA, NITIN. **TMSK: text-miner software Kit**. 2004

JACOB, J. P. **Somos todos informívoros**. INFO EXAME, São Paulo, v. 20, n. 233, pág. 26, 2005.

LE CODIAC, YVES-FRANÇOIS. **A Ciência da Informação**. Brasília, DF : Brique de Lemos, 2004.

LOH, STANLEY. Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos. Porto Alegre : UFRGS, 2001. 110p. Tese (Ciência da Computação) – Programa de Pós-graduação em Computação da Universidade Federal do Rio Grande do Sul.

KONCHADY, Manu. Text mining application programming. Massachusetts : Charles River Media, 2006.

LOPES, M. C. S., 2004, Mineração de dados textuais utilizando técnicas de clustering para o idioma português. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MANNING, Christopher D; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. An introduction to Information Retrieval. Cambridge : Cambridge University Press, 2005 [Preliminary Draft].

MARCONDES, C H. Metadados: descrição e indexação de recursos informacionais na Web. [Apresentações de aula]. 2003.

MATSUNAGA, LILIAM AYAKO. Uma Metodologia de Categorização Automática de Textos para a Distribuição dos Projetos de Lei às Comissões Permanentes da Câmara Legislativa do Distrito Federal [Rio de Janeiro] 2007

MEADOWS, A. J. **A comunicação científica**. Brasília: Brique de Lemos, 2000

ORENGO, V.M.; HUYCK, C. A stemming algorithm for the portuguese language. In: String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on , vol., no., pp. 186-193, 13-15 Nov. 2001. Disponível em:<

[http://ieeexplore.ieee.org/iel5/7760/21327/00989755.pdf?isnumber=21327\[\]=STD&arnumber=989755&arnumber=989755&arSt=+186&ared=+193&arAuthor=+Orengo%2C+V.M.%3B++Huyck%2C+C](http://ieeexplore.ieee.org/iel5/7760/21327/00989755.pdf?isnumber=21327[]=STD&arnumber=989755&arnumber=989755&arSt=+186&ared=+193&arAuthor=+Orengo%2C+V.M.%3B++Huyck%2C+C). Acesso em: 04 nov. 2007.

PEREIRA, Renato Milagres, SANTOS, Luis Filipe de Mello. Ferramenta Midas-UFF: módulo de classificação. Niterói : UFF, 2004, 65p. Monografia (Ciência da Computação) - Departamento de Ciência da Computação da Universidade Federal Fluminense.

SEBASTIANI, Fabrizio. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, v. 34, n.1, Mar. 2002.

TAN, AH-HWEE. **Text Mining: the state of the art and the challenges**. 1999. Disponível em: <http://www.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf>. Acesso: 20 abr. 2006.

WIVES, LEANDRO KRUG. Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering". Porto Alegre : UFRGS, 1999. 102p. Dissertação (Ciência da Computação) - Programa de Pós-graduação em Computação da Universidade Federal do Rio Grande do Sul.

YANG, Y.; LIU, X. **A re-examination of text categorization methods**. In The 22th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pages 42-49, 1999.

YANG, YIMING; PEDERSEN, JAN O. A Comparative Study on Feature Selection in Text Categorization. 1997. Disponível em: <<http://online.mq.edu.au/pub/COMP348/resources/yangpedersen97.pdf>>. Acesso em: 30 jan 2007.

WEISS, SHOLOM et al. Text mining: predictive methods for analyzing unstructured Information. New York: Springer, 2004.

APÊNDICE 1

LISTAGEM DOS ARQUIVOS DO CORPUS EM PORTUGUÊS

Tamanho	Nome
135.687	DM-A Inteligencia Competitiva modelando o Sistema de Informacao de Clientes - Finep.txt
34.808	DM-A tecnologia como suporte a inteligencia competitiva.txt
105.514	DM-Desenvolvimento de Metodologia para a Previsao de Consumo de Energia Eletrica Utilizando Redes Neurais.txt
9.213	DM-EXPERT SYSTEMS FEITOS PELA EQUIPE NELSON.txt
31.778	DM-INTELIGENCIA COMPETITIVA - uma abordagem sobre a coleta de informacoes publicadas.txt
29.716	DM-INTELIGENCIA COMPETITIVA E A GESTAO DO CONHECIMENTO.txt
6.553	EFEITOESTUFA-RIO 3- Evento Mundial sobre Clima e Energia - LAREF 2003 - Feira Tecnologica Latino Americana de Energias Renovaveis.txt
5.221	EFICIENCIA-ACOMPANHAMENTO DO CONSUMO DE ENERGIA ELETTRICA NA PRESIDENCIA DA REPUBLICA EM DEZEMBRO DE 2002 - GRAFICO.txt
4.974	EFICIENCIA-ANEEL - RESOLUCAO N 334, DE 2 DE DEZEMBRO DE 1999 - Autoriza as concessionarias de servico publico de energia eletrica a desenvolverem projetos visando a melhoria do fator de carga.txt
26.588	EFICIENCIA-Bases Conceituais do projeto casa autonoma.txt
22.058	EFICIENCIA-Compensacao Reativa - INEPAR [FOLDER].txt
30.128	EFICIENCIA-Correntes na Ligacao de Bancos Capacitadores [FOLDER].txt
5.167	EFICIENCIA-FLOSTAR-M_99=04 - Schlumberger PT-BR[FOLDER].txt
33.945	EFICIENCIA-Guia de Aplicacao de Capacitores BT - FOLDER - INEPAR.txt
214.779	EFICIENCIA-MME - IMPLEMENTACAO DA LEI DE EFICIENCIA ENERGETICA - CGIEE - Relatorio de atividades - maio a dezembro de 2002.txt
69.578	EFICIENCIA-Plano de trabalho - Implementacao da Lei de Eficiencia Energetica - COMITE GESTOR DE INDICADORES E NIVEIS DE EFICIENCIA ENERGETICA - CGIEE.txt
22.253	EFICIENCIA-PROCEL - CADASTRO DE PREDIOS PUBLICOS FEDERAIS - MANUAL DE PREENCHIMENTO.txt
6.343	EFICIENCIA-Programas de Eficiencia Energetica - Ciclo 1999-2000.txt
58.997	ENERGIA-A Industria de Energia Eletrica Brasileira - Reestruturacao, Competicao e Contestabilidade.txt
47.763	ENERGIA-A Questao do Investimento no Setor Eletrico Brasileiro - Reforma e Crise.txt
5.481	ENERGIA-ANEEL - RESERVA DE CAPACIDADE AUTOPRODUTOR - res1999371.txt
27.662	ENERGIA-ANEEL - RESOLUCAO N 223, DE 29 DE ABRIL DE 2003 - condicoes gerais para elaboracao dos Planos de Universalizacao de Energia Eletrica.txt
15.334	ENERGIA-ANEEL - RESOLUCAO N 278, DE 19 DE JULHO DE 2000 - Limites e condicoes para participacao dos agentes economicos nas atividades do setor de energia eletrica.txt
157.981	ENERGIA-ANEEL - RESOLUCAO N. 456, DE 29 DE NOVEMBRO DE 2000 - Condicoes Gerais de Fornecimento de Energia Eletrica.txt
67.418	ENERGIA-Diagnostico da Crise de Oferta de Energia Eletrica Contribuicoes para o Equacionamento - Estudo Preliminar - FIERN.txt
177.558	ENERGIA-Formacao de Precos e Comercializacao de Energia no Novo Ambiente Institucional do Sistema Eletrico Brasileiro - GERENCIAMENTO DE RISCO NO SETOR DE ENERGIA ELETTRICA.txt

9.920 ENERGIA-FORUM SUDENE ENERGIA - A construcao de um modelo [FORUM].txt
97.684 ENERGIA-Impactos Fiscais da Crise de Energia Eletrica - 2001 e 2002.txt
39.614 ENERGIA-LUIS EDUARDO MAGALHAES - 3 - ENERGIA - NOVOS CENARIOS - Universalizacao do acesso, uso racional e fontes alternativas para o futuro.txt
478.450 ENERGIA-MINISTERIO DAS MINAS E ENERGIA - ACOES DO ANO 2000 DO MME.txt
518.750 ENERGIA-MME - Camara de Gestao do Setor Eletrico - Comite de Revitalizacao do Modelo do Setor Eletrico - GT6 - Expansao do sistema de transmissao.txt
156.789 ENERGIA-MME - Proposta de Modelo Institucional do Setor Eletrico.txt
162.200 ENERGIA-MME - CNPE - CGSE - COMITE DE REVITALIZACAO DO MODELO DO SETOR ELETRICO - Relatorio de Progresso N 4 - ANO 2002.txt
6.064 ENERGIA-Nao existe energia limpa.txt
47.653 ENERGIA-NI 273 anexo 1 - Discurso do senador - Senador Delcideo Amaral (PT - MS) - Carlos Nascimento.txt
26.786 ENERGIA-Nota Tecnica no 30-2003-SEM-ANEEL - Alteracao da sistemática de estabelecimento do preco minimo do Mercado de Curto Prazo.txt
29.910 ENERGIA-Ref PROPOSTAS SOBRE A AUDIENCIA PUBLICA AP005-2002.txt
149.005 ENERGIA-Relatorio da Comissao de Analise do Sistema Hidrotermico de Energia Eletrica.txt
44.234 ENERGIA-RESPOSTAS DE CURTO PRAZO PARA A CRISE DE ENERGIA ELETRICA A PROPOSTA DE UM PROGRAMA DE GERACAO DISTRIBUIDA - PROGEDIS.txt
5.799 ENERGIA-Sustentabilidade na geracao e uso de energia no Brasil - os proximos 20 anos.txt
40.539 EOLICA-ASPECTOS PRINCIPAIS RELACIONADOS A IMPLANTACAO DE PARQUES EOLICOS NO SISTEMA ELETRICO.txt
5.848 EOLICA-Assunto - Orientacao da Rosa dos Ventos no SISLEV.txt
37.170 EOLICA-OTIMIZACAO DOS PARAMETROS DA DISTRIBUICAO DE WEIBULL.txt
12.488 EOLICA-Problemas na Estacao 2101 - Salina de Massambaba - Torre de Referencia - Set2002.txt
160.596 EOLICA-Projeto Final - ENERGIA EOLICA (parte1).txt
8.637 EOLICA-Projeto Final - ENERGIA EoLICA - CAPITULO 9 - PLANILHA DE APOIO A ESCOLHA DO AEROGERADOR E REFERENCIAS BIBLIOGRAFICAS - (parte2).txt
7.812 EOLICA-Projeto Final - ENERGIA EoLICA - FLUXOGRAMA DO CALCULO DA ENERGIA PRODUZIDA - (parte1-energia) DIAGRAMA.txt
2.826 EOLICA-Rosa dos ventos(exemplo).txt
29.321 EOLICA-Vaisala MAWS301 - Estacao meteorologica automatica [FOLDER].txt
37.842 EOLICA-Valor Economico da Tecnologia Especifica da Fonte - VETEF - Programa de Incentivo as Fontes Alternativas de Energia Eletrica - PROINFA (1a Etapa).txt
3.883 EQUIPELETRICO-Auto Transformador Compensadores de Partida para Instalacao Abridada Modelo ATC [FOLDER].txt
2.592 EQUIPELETRICO-DADOS TECNICOS TRANSFORMADORES [FOLDER].txt
2.966 EQUIPELETRICO-Desenhos com modelos de transformadores.txt
6.805 EQUIPELETRICO-Reatores a Nucleo de Ar FOLDER.txt
5.479 EQUIPELETRICO-Reatores para Limitacao de Correntes de Carga Bancos de Capacitores FOLDER.txt
9.788 EQUIPELETRICO-Reatores Trifasicos de Linha FOLDER.txt
4.608 EQUIPELETRICO-Transformador de Potencia a Seco, Moldado em Resina Epoxi FOLDER.txt
3.165 EQUIPELETRICO-Transformadores Monofasicos ate 1500 VA Modelo TMH FOLDER.txt
3.484 EQUIPELETRICO-Transformadores Monofasicos at, 1500 VA FOLDER.txt

4.048 EQUIPELETRICO-Transformadores Monofasicos at, 25 KVA Modelo TMA - FOLDER.txt

3.292 EQUIPELETRICO-Transformadores Monofasicos com Caixa FOLDER.txt

3.666 EQUIPELETRICO-Transformadores Trifasicos com Caixa folder.txt

4.370 EQUIPELETRICO-Transformadores Trifasicos de Baixa Tensao para Instalacao Abridada Modelo TRT FOLDER.txt

7.624 EQUIPELETRICO-Transformadores Trifasicos de Media Tensao para Instalacao Abridada Modelo TRT-MT FOLDER.txt

1.230 EVASAO-Curso de Engenharia Eletrica - Calculo, Algebra e Fisica Turmas de 1-99 a 2-00 - Grafico II - Aprovacao em 2 semestres.txt

1.109 EVASAO-Curso de Engenharia Eletrica - Calculo, Algebra e Fisica Turmas de 1-99 a 2-00 -Grafico I - Aprovacao no 1 semestre.txt

31.649 GD-ENERGIA DA CANA DE ACUCAR NO BRASIL.txt

34.479 GD-Experiencia em plantas de co-geracao de pequeno porte.txt

76.634 GD-NOTAS SOBRE GERACAO DISTRIBUIDA - Introducao - INEE.txt

32.947 GERACAO-Presidencia - LEI No 9.074, DE 7 DE JULHO DE 1995 - Estabelece normas para outorga e prorrogacoes das concessoes e permissoes de servicos publicos.txt

15.676 GERENCIAMENTO-Gerenciando Projetos na Era de Grandes Mudancas - Uma breve abordagem do panorama atual.txt

6.078 GERENCIAMENTO-GOVERNO ELETRONICO E GESTAO DO CONHECIMENTO.txt

13.626 GERENCIAMENTO-Modelo de Maturidade Organizacional de Gerencia de Projetos Organizational Project Management Maturity Model - OPM3 - Um trabalho voluntario.txt

72.320 GERENCIAMENTO-RECURSOS HUMANOS EM PEQUENAS EMPRESAS.txt

3.924 LABLUX-Laboratorio de Luminotecnica - MA-8.txt

46.910 MANUTENCAO-IFS Manutencao - FOLDER.txt

11.624 MANUTENCAO-MAXIMO Extended Enterprise - A solucao para gestao do processo de manutencao que faz de tudo do jeito que voc` quiser SOFTWARE FOLDER.txt

356.715 MUDANCACLIM-Agenda 21 Brasileira - ACOES PRIORITARIAS.txt

218.756 OPERACAO-ONS - Operacao do Sistema Interligado Nacional - Dados relevantes de 2000.txt

128.972 OPERACAO-ONS - Operacao do Sistema Interligado Nacional - Dados relevantes de 2001.txt

5.831 PED-ANEEL - APLICACAO DE RECURSOS COMBATE AO DESPERDICIO - res2000271.txt

27.141 PED-ANEEL - FORMULARIO DE PROJETO - Modelo Matematico - Previsao de Vazoes - Controle de Cheias - SIMULA.txt

39.194 PED-ANEEL - Formulário de Projeto - Aplicacao de Tecnicas de Data Mining para determinacao de indices de demanda para Unidades Consumidoras n/Eo residenciais de Baixa Tensao - DEMANDA.txt

38.934 PED-ANEEL - Formulário de Projeto - Aplicacao de Tecnicas de Data Mining para estabelecimento de curva de correlacao do Consumo kWh x Demanda kVA - CORREL.txt

27.294 PED-ANEEL - FORMULARIO DE PROJETO - Construcão de Modelo de Simulacao de Vazoes da Bacia do Rio Paraiba do Sul Utilizando Data Mining.txt

24.196 PED-ANEEL - Formulário de Projeto - Data Mining para Gestao Energetica no Segmento de Grandes Clientes.txt

26.463 PED-ANEEL - Formulário de Projeto - Data Mining para Marketing Inteligente no Segmento de Grandes Clientes.txt

66.692 PED-ANEEL - FORMULARIO DE PROJETO - DESENVOLVIMENTO DE SOFTWARE INTELIGENTE PARA ANALISE DE FRAUDE EM CONSUMIDORES.txt

39.529 PED-ANEEL - Formulário de Projeto - DESENVOLVIMENTO DE SOFTWARE PARA ANALISE ECONOMICA DA UTILIZACAO DE GERACAO DISTRIBUIDA PARA SUPRIMENTO DE CONSUMIDORES NAO ATENDIDOS - GERDIS.txt

57.070 PED-ANEEL - Formulário de Projeto - GESTAO ENERGETICA PARA GRANDES CLIENTES - GESTENERG.txt

18.472 PED-ANEEL - Formulário de Projeto - Introducao de Inibidores de Furto (INIFUR) em Alimentadores Monofasicos e Trifasicos, A, reos e Subterraneos de BT.txt

40.689 PED-ANEEL - Formulário de Projeto - Introducao de Inibidores de Furto em Alimentadores Trifasicos, Aereos e Subterraneos de BT - INIFUR.txt

58.230 PED-ANEEL - Formulário de Projeto - INTRODUÇÃO NA MATRIZ ENERGÉTICA BRASILEIRA DE PCH'S DE BAIXA QUEDA PELA UTILIZAÇÃO GERADORES ASSÍNCRONOS.TXT

39.395 PED-ANEEL - FORMULÁRIO DE PROJETO - Reator Linear Controlado do Tipo Transformador - RLCTT.txt

48.169 PED-ANEEL - Formulário de Projeto - RESGATE ENERGÉTICO - AUMENTO DE RENDIMENTO NA PRODUÇÃO DE ENERGIA EM USINAS EXISTENTES.txt

33.143 PED-ANEEL - Formulário de Projeto - RESGATE ENERGÉTICO - AUMENTO DE RENDIMENTO NA PRODUÇÃO DE ENERGIA EM USINAS.txt

46.647 PED-ANEEL - FORMULÁRIO DE PROJETO - Sistema de Inibição de Fraudes no Sistema de Média e Baixa Tensão da LIGHT.txt

38.194 PED-ANEEL - Formulário de Projeto - Sistema de lacre com monitoramento de violação micro-eletrônica wireless - IMBELTRANS.txt

55.742 PED-ANEEL - Formulário de Projeto - Sistema de Medição Desarticulador de Fraudes - DESART.txt

38.072 PED-ANEEL - Formulário de Projeto - Sistema inteligente de medição e distribuição anti-fraudes - IMBELMED.txt

44.266 PED-ANEEL - FORMULÁRIO DE PROJETO - SISTEMA PARA IDENTIFICAÇÃO E ATRACÃO DE CONSUMIDORES LIVRES.txt

30.655 PED-ANEEL - Formulário de Projeto - TÉCNICAS DE DATA MINING PARA IDENTIFICAÇÃO DE LOCAIS PARA PARQUES EÓLICOS EM SISTEMAS GEOGRÁFICOS DE INFORMAÇÃO.txt

40.055 PED-ANEEL - LIGHT - SERVIÇOS E ELETRICIDADE SA [Resumo dos Projetos].txt

114.296 PED-ANEEL - Manual dos programas de pesquisa e desenvolvimento tecnológico do setor elétrico brasileiro.txt

578.508 PED-ANEEL - MANUAL PARA ELABORAÇÃO DO PROGRAMA ANUAL DE COMBATE AO DESPERDÍCIO DE ENERGIA ELÉTRICA.txt

33.541 PED-ANEEL - PROJETO - Sistema de Aferição de Medidores exteriorizados tipo CP-Redes - AFERI.txt

16.417 PED-CARTA-CONVITE AS EMPRESAS BRASILEIRAS - FUNDO DE ESTÍMULO A INTERAÇÃO UNIVERSIDADE EMPRESA - FUNDO VERDE-AMARELO.txt

44.321 PED-Ct-energia - Propostas de investimento do país em linhas de P_e_D para a área de geração de energia elétrica.txt

12.548 PED-Estado da Arte da Qualidade da Energia Elétrica - Workshop.txt

55.700 PED-ESTUDO DE CASO - FABRICA DO AGRICULTOR -- DO ESTADO DO PARANÁ, SUL DO BRASIL.txt

26.804 PED-FINEP -INOVA, BRASIL - ACOES EM ANDAMENTO DOS FUNDOS SETORIAIS - Carta_convite_empresa_fva_anexo.txt

24.846 PED-FUNDOS SETORIAIS - DESAFIOS PARA CIÊNCIA E TECNOLOGIA NO CONTEXTO DO SETOR ELÉTRICO.txt

35.117 PED-MCT - Secretaria Executiva - Assessoria de Acompanhamento e Avaliação - Plano Plurianual - PPA 2004-2007 - PROGRAMAS E ACOES - PROPOSTA QUALITATIVA.txt

4.345 PED-PROJETO DE P_E_D DA UFF - RESGATE ENERGÉTICO - DESCRIÇÃO DAS ETAPAS.txt

46.983 PED-PROPOSTA DE ÁREAS RELEVANTES PARA ATIVIDADES DE P_E_D - CT ENERGIA-COPPE.txt

18.240 RENOVÁVEIS-ANEEL - RESOLUÇÃO N 488, DE 29 DE AGOSTO DE 2002 - Resolução sobre o valor normativo.txt

51.278 RENOVÁVEIS-DIRETIVA 2001-77-CE DO PARLAMENTO EUROPEU E DO CONSELHO relativa a promoção da eletricidade produzida a partir de fontes de energia renováveis no mercado interno de eletricidade.txt

773.751 RENOVÁVEIS-INSERÇÃO DE CENTRAIS COGERADORAS A BAGACO DE CANA.TXT

40.343 RENOVÁVEIS-O Contexto das Energias Renováveis no Brasil.txt

23.966 RENOVÁVEIS-Tecnologia Solar no Brasil - Os próximos 20 anos.txt

2.150 SIFF-Carta SIFF rescisão mma - Rescisão do Contrato de Prestação de Serviços celebrado em 01 de setembro de 2001.txt

11.349 SIFF-Curso Energia Eólica - DEWI-Brasil [FOLDER].txt

15.024 SIFF-Estação 2101 - Salina de Massambaba - Torre de Referência - Cabo Frio, Brasil - Dezembro 2002.txt

13.718 SIFF-Estação 2103 - Praia de Pernambuco - Torre Satélite 2 - Cabo Frio, Brasil - Fevereiro de 2002.txt

4.559 SIFF-Estação 2101 - Ponta de Massambaba - Torre de Referência - Cabo Frio, Brasil - Abril de 2002.txt

4.698 SIIF-Estacao 2101 - Ponta de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Maio de 2002.txt
4.402 SIIF-Estacao 2101 - Ponta de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Marco de 2002.txt
4.511 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Dezembro de 2001.txt
4.682 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Novembro de 2001.txt
5.266 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Outubro de 2001.txt
5.378 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Setembro de 2001.txt
27.174 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Dezembro de 2001.txt
18.638 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Janeiro de 2002.txt
15.315 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Novembro 2002.txt
19.283 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Marco de 2002.txt
15.743 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Abril de 2002.txt
15.056 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Agosto de 2002.txt
18.893 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Fevereiro de 2002.txt
15.614 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Julho de 2002.txt
15.891 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Junho de 2002.txt
15.680 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Maio de 2002.txt
25.677 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Novembro de 2001.txt
15.317 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Novembro de 2002.txt
26.127 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Outubro de 2001.txt
14.987 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Outubro de 2002.txt
22.861 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Setembro de 2001.txt
14.802 SIIF-Estacao 2101 - Salina de Massambaba - Torre de Referencia - Cabo Frio, Brasil - Setembro de 2002.txt
4.584 SIIF-Estacao 2102 - Ponta de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Dezembro de 2001.txt
4.652 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Abril de 2002.txt
4.648 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Marco de 2002.txt
11.275 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Outubro de 2002.txt
4.831 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Maio de 2002.txt
4.135 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Outubro de 2001.txt
4.183 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Setembro de 2001.txt
4.182 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 2 - Cabo Frio, Brasil - Setembro de 2001.txt
11.360 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Agosto de 2002.txt
13.283 SIIF-Estacao 2102 - Salina de Massambaba - Torre Satelite 1 - Cabo Frio, Brasil - Marco de 2002.txt

APÊNDICE 2

LISTAGEM DOS ARQUIVOS DO CORPUS EM INGLÊS

Tamanho	Nome
62.259	AUTOMACAO-Intelligent applications for the management of electrical systems in industrial plants.txt
141.689	CONCESSIONARIA-REVIEW OF CAPITAL EXPENDITUR REQUIREMENTS-QUEENSLAND.txt
141.702	CONCESSIONARIA-POWERLINK QUEENSLAND - Review of Capital Expenditure Requirements.txt
94.139	DADOS-SB110 REPORT - The California Energy Commission's Reporting, Forecasting & Data Collection Responsibilities.txt
147.812	DISTRIBUICAO-Complex thermal applications - Easy measurements - Portable, modular flue gas analyser for industrial processes - Folder TESTO.txt
36.329	DISTRIBUICAO-Knowledge-Based System for Distribution System Outage Locating Using Comprehensive Information.txt
38.846	DM-A Fuzzy Expert System for the Integrated Fault Diagnosis.txt
2.016	DM-A Hybrid Fault Diagnosis System Using ANNs-ES Techniques - ABSTRACT.txt
32.004	DM-A PARTICLE SWARM OPTIMIZATION FOR REACTIVE POWER AND VOLTAGE CONTROL CONSIDERING VOLTAGE STABILITY.txt
7.529	DM-Data Analyst- SILICON [FOLDER].txt
183.274	DM-DATA MINING - AN INTRODUCTION.txt
8.775	DM-Energy Data Warehouse - FLEXIBLE DATA COLLECTION, DISTRIBUTION AND ANALYSIS OF METERED CONSUMPTION MEETS DYNAMIC MARKET NEEDS [Silicon Energy].txt
132.013	DM-Industrial use of safety-related artificial neural networks.txt
32.093	EDUCACAO-A PROBLEM BASED LEARNING APPROACH FOR FRESHMAN ENGINEERING.txt
35.593	EDUCACAO-APPLICATION OF EDUCATIONAL AND ENGINEERING RESEARCH TO CLASSROOM TEACHING.txt
32.556	EDUCACAO-Assessing Diverse Student Outcomes More Efficiently - An Example From Engineering Education.txt
25.323	EDUCACAO-Canadian Conference on Engineering Education Program - CONFERENCE.txt
10.363	EDUCACAO-CONCEPT MAPS AND COMPETENCES CHARTS-A ROADMAP.txt
54.492	EDUCACAO-CONCEPT MAPS FOR ENGINEERING EDUCATION.txt
8.767	EDUCACAO-COURSE OUTLINE 2000-2001 - Engineering 2C03 - Electricity, Thermophysics and Energy.txt
41.296	EDUCACAO-Curriculum Integration - Students Linking Ideas across Disciplines.txt
40.994	EDUCACAO-INTEGRATING KNOWLEDGE ACROSS THE ENGINEERING CURRICULUM.txt
34.014	EDUCACAO-KNOWLEDGE CONSTRUCTION AND SHARING IN QUORUM.txt
35.172	EDUCACAO-Knowledge Maps for Intelligent Questioning Systems in Engineering Education.txt
29.583	EDUCACAO-Nonlinear Concepts and Methods - Non Linear Systems, Learning Systems - Lecture 2.txt
16.850	EDUCACAO-PROGRESS REPORT. APE TRACK A.txt
26.194	EDUCACAO-SmartTutor - Combining SmartBooks and Peer Tutors for Multi-media On-Line Instruction.txt

207.607 EDUCACAO-STRATEGIC PLAN - THE UNIVERSITY OF WESTERN AUSTRALIA.txt

20.784 EDUCACAO-TALLOIRES DECLARATION INSTITUTIONAL SIGNATORY LIST.txt

86.310 EDUCACAO-TALLOIRES DECLARATION RESOURCE KIT.txt

28.625 EDUCACAO-The Cognitive Flexibility Theory - an Approach for Teaching Hypermedia Engineering.txt

28.322 EDUCACAO-THE GIANT-KNOWLEDGE CONSTRUCTION & SHARING.txt

71.323 EDUCACAO-UNIVERSITY OF WESTERN AUSTRALIA - ACHIEVING INTERNATIONAL EXCELLENCE - An Operational Priorities Plan for 2003-2005.txt

38.496 EDUCACAO-USING CONCEPT MAPPING AS AN INSTRUCTIONAL STRATEGY.txt

34.897 EDUCACAO-USING CONCEPT MAPS AND CONCEPT QUESTIONS TO ENHANCE CONCEPTUAL UNDERSTANDING.txt

4.570 EFICIENCIA-A Green Building Forum for Young Professionals USGBC Green Building International Conference and Expo.txt

7.929 EFICIENCIA-ALARM MANAGER - DISCOVER INEFFICIENCIES IN ENERGY CONSUMING EVENTS AS THEY OCCUR.txt

175.657 EFICIENCIA-ALTERNATIVES TO COMPRESSOR COOLING - PIER - 2002-01-10_600-00-003.txt

21.401 EFICIENCIA-Better Buildings by Design - We shape our buildings - thereafter they shape us.txt

413.746 EFICIENCIA-BUILDING ENERGY MANAGEMENT SYSTEMS - APPENDIX I - SITUATION ASSESSMENT - TECHNOLOGY SCANNING.txt

540.201 EFICIENCIA-Cooling System Selection - CSW CORPORATION.txt

9.412 EFICIENCIA-Cost Analyst - ENTERPRISE-LEVEL ENERGY BILL COST ANALYSIS AND ALLOCATION FOR GAS, ELECTRICITY AND OTHER COMMODITIES [FOLDER].txt

16.813 EFICIENCIA-Department of Environmental Protection - Cambria Office Building - Highlighting - high performance.txt

83.402 EFICIENCIA-ENERGY ACCOUNTING - A Key Tool in Managing Energy Costs HANDBOOK.txt

340.254 EFICIENCIA-ENERGY DESIGN GUIDELINES FOR HIGH PERFORMANCE SCHOOLS - HOT AND HUMID CLIMATES.txt

34.260 EFICIENCIA-Enterprise Energy Management Software [FOLDER].txt

6.788 EFICIENCIA-Enterprise Navigator [Silicon Energy].txt

8.294 EFICIENCIA-ESPC Contract Risk - Responsibility Implications.txt

30.393 EFICIENCIA-EUCI Presents - Energy Risk Management Week.txt

72.101 EFICIENCIA-Federal Energy Management Program M&V Guidelines - Measurement and Verification for Federal Energy Projects - Renewable Energy Technologies - Chapter 35.txt

202.228 EFICIENCIA-GREEN CRITERIA - BC HYDRO.txt

557.541 EFICIENCIA-GREENING FEDERAL FACILITIES - An Energy, Environmental and Economic Resource Guide for Federal Facility Managers.txt

992.690 EFICIENCIA-GREENING FEDERAL FACILITIES - An Energy, Environmental, and Economic Resource Guide for Federal Facility Managers and Designers - SECOND EDITION.txt

559 EFICIENCIA-HIGH PERF GUIDELINES - COVER.txt

56.334 EFICIENCIA-HIGH PERF GUIDELINES - INTRODUCTION.txt

15.544 EFICIENCIA-HIGH PERF GUIDELINES - PREFACE.txt

21.558 EFICIENCIA-HIGH PERFORMANCE - ACKNOWLEDGMENTS.txt

84.608 EFICIENCIA-High Performance Guidelines - Triangle Region Public Facilities - APPENDIX.txt

58.826 EFICIENCIA-High Visibility Energy and Power Quality Compliance Meters - FOLDER.txt

103.931 EFICIENCIA-HIGH-PERFORMANCE COMMERCIAL BUILDINGS - A TECHNOLOGY ROAD MAP.txt

152.200 EFICIENCIA-HOW TO HIRE AN ENERGY AUDITOR - To Identify Energy Efficiency Projects HANDBOOK.txt

110.400 EFICIENCIA-HVAC&R RESEARCH FOR THE 21 CENTURY.txt

81.997 EFICIENCIA-Initial Applications of the FEMP Measurement & Verification Guidelines in Super ESPC Delivery Orders - Observations & Recommendations.txt

13.154 EFICIENCIA-KILO-EEM KIT - ENERGY SUPERVISION AND CONTROL KIT [FOLDER].txt

10.646 EFICIENCIA-KILO-SF Three-phase electrical energy analyser [FOLDER].txt

979.056 EFICIENCIA-M&V GUIDELINES - MEASUREMENT AND VERIFICATION FOR FEDERAL ENERGY PROJECTS-26265.txt

69.844 EFICIENCIA-M&V SAVINGS FROM IMP IN OPER&MAINT OR ENERGY CONSUMING SYSTEM - OMpaper.txt

13.673 EFICIENCIA-M&v SAVINGS FROM OPERATION AND MAINTENANCE IMPROVEMENTS - tech_note_2.txt

.093.197 EFICIENCIA-METHOD FOR DETERMINING MEASUREMENT ACCURACY AND DATA STORAGE FREQUENCY FOR IMPROVED BUILDING ENERGY EFFICIENCY.txt

10.288 EFICIENCIA-Office of Power Technologies - Desiccant technology keeps indoor air fresh, healthy, comfortable.txt

402.904 EFICIENCIA-Operational Guidelines for Baseline Studies, Validation, Monitoring and Verification of Joint Implementation Projects.TXT

8.186 EFICIENCIA-Security in Sensor Networks - Proving location claims.txt

4.646 EFICIENCIA-Silicon Energy - Energy Analyst - Software Description.txt

24.789 EFICIENCIA-Silicon Energy - INTEGRATE ENERGY SUPPLY AND DEMAND - Software Description.txt

13.370 EFICIENCIA-State-of-the-Art Building Concepts Lower Energy Bills.txt

14.470 EFICIENCIA-SUSTAINABLE BUILDING MATERIALS & SYSTEMS SHOW CLIMATE CHANGE.txt

44.455 EFICIENCIA-Tools for Sustainable Building.txt

14.658 EFICIENCIA-VisIR - Infrared Vision Cameras - Folder.txt

34.290 ENERGDISTRIB-AN EVOLUTIONARY APPROACH FOR CAPACITOR PLACEMENT IN DISTRIBUTION NETWORKS.txt

28.371 ENERGDISTRIB-Distributed Energy & Electric Reliability - Fact Sheet - Micro-sources with Storage Bringing High Value to Customers.txt

23.717 ENERGDISTRIB-DISTRIBUTED ENERGY RESOURCES - AN UPDATE.txt

9.965 ENERGDISTRIB-Distributed Energy Resources Program.txt

12.469 ENERGDISTRIB-DISTRIBUTED ENERGY RESOURCES- NATIONAL PERSPECTIVE.txt

260.091 ENERGDISTRIB-DISTRIBUTED GENERATION INTERCONNECTION RULES - Supplemental Recommendation Regarding.txt

126.127 ENERGDISTRIB-Distributed Generation Strategic Plan - CALIFORNIA ENERGY COMMISSION.txt

39.515 ENERGIA-ECONOMIC DEVELOPMENT & ENERGY COMMITTEE OF THE SUFFOLK COUNTY LEGISLATURE - Minutes.txt

179.215 ENERGIA-ELECTRANEST SA CAPITAL EXPENDITURE REVIEW - Australian Competition and Consumer Commission.txt

179.165 ENERGIA-ElectraNet SA Capital Expenditure Review.txt

5.237 ENERGIA-ELECTRICITY MARKET - CONCEPTUAL DESIGN.txt

137.938 ENERGIA-Electricity Market - Model Analysis.txt

21.362 ENERGIA-EVOLVING ENERGY ENTERPRISE - POSSIBLE ROAD AHEAD R&D GRAND CHALLENGES.txt

182.136 ENERGIA-Identification of Issues for the Development of Regional Power Markets in South America.txt

3.123 ENERGIA-Improving Mid-Term Energy Forecasts for Buildings.txt

5.857 ENGENHARIA-McMASTER UNIVERSITY FINAL EXAMINATION - ENGINEERING 2C03 [PROVA DE ABRIL].txt

270.850 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Affected Environment - CHAPTER 3.txt

8.394 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix A.txt

10.341 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix B.txt

12.961 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix C.txt

39.321 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix D.txt

50.349 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix E.txt

9.125 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix F.txt

189.471 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix G.txt

1.058 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Appendix H.txt

41.589 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Description of Alternatives - CHAPTER 2.txt

214.443 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Environmental Consequences - CHAPTER 4.txt

369 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - FINAL - Environmental Assessment [SOMENTE CAPA] - .txt

10.695 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - List of Preparers - CHAPTER 5.txt

7.274 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - Purpose of and Need for the Proposed Action - CHAPTER 1.txt

17.202 EOLICA-20-MW Windfarm and Associated Energy Storage Facility - References - CHAPTER 6.txt

202.046 EOLICA-A Feasibility Study to Develop Local and Regional Use of Wind Energy on the Kola Peninsula, Murmansk Region, Russia.txt

9.412 EOLICA-A global strategy for Wind Energy - The substitution of all nuclear and fossil energies by renewable energies.txt

22.175 EOLICA-A PRELIMINARY INVESTIGATION OF TWO SMALL-SCALE, AUTONOMOUS WIND-HYDROGEN SYSTEMS.txt

103.761 EOLICA-A Summary of Environmentally Friendly Turbine Design Concepts.txt

10.672 EOLICA-Address of Dr. Hermann Scheer, German MP and General Chairman of the World Council for Renewable Energy, to the Global Windpower Conference 2002 in Paris.txt

2.839 EOLICA-Appointment of Consultants to Investigate the Impact of Increased Levels of Wind Penetration on the Electricity Systems in the Republic of Ireland and Northern Ireland - CER.txt

303.262 EOLICA-Assessing the Economic Development - Impacts of Wind Power - FINAL REPORT.txt

6.432 EOLICA-Brazil Ceara coast wind & production - Energy calculation.txt

24.966 EOLICA-Budget offer - Sondot,cnica.txt

2.442 EOLICA-Case Studies of Development Approaches and Opportunities for Wind Power in Wisconsin.txt

30.939 EOLICA-Cover Sheet - Environmental Assessment - 20-MW Windfarm and Associated Energy Storage Facility.txt

8.086 EOLICA-DRAFT NWCC Guidelines for Assessing the Economic Development Impacts of Wind Power.txt

5.866 EOLICA-European wind industry - another record year - ?5.8 billion European market in 2002.txt

66.821 EOLICA-General Specification - V66 - 1.75 MW - OptiSpeedTM - Wind Turbine.txt

74.439 EOLICA-General Specification V52 - 850 kW OptiSpeedTM - Wind Turbine.txt

38.484 EOLICA-General Specification V66 - 2.0 MW offshore OptiSpeedTM - Wind Turbine.txt

37.158 EOLICA-General Specification V80 - 2.0 MW offshore OptiSpeedTM - Wind Turbine.txt

23.109 EOLICA-Global Wind Energy Market Report - Wind Energy Turns in Strong Performance in 2001 New Records Set in Europe, United States, India.txt

583.449 EOLICA-Guided Tour on Wind Energy.txt

10.056 EOLICA-Guidelines for Assessing the Economic Development Impacts of Wind Power.txt

82.274 EOLICA-IEA-IMPLEMENTING AGREEMENT RESEARCH WIND TURBINE - EXECUTIVE COMMITTEE.txt

127.842 EOLICA-INTERACTIONS - WIND FARMS - BULK POWER - MARKETS.txt

9.194 EOLICA-INVESTING IN WIND POWER - How To Put Your Money Where Your Mouth Is.txt

4.673 EOLICA-New hope for the Danish wind market after the 2001 break-down - Denmark distances from certificate system.txt

2.736 EOLICA-NORDEX N-60 Annual energy yield [FOLDER].txt

15.589 EOLICA-NORDEX N-60 Electrical installation [FOLDER].txt

4.156 EOLICA-NORDEX N-60 Lightning and overvoltage protection.txt

2.256 EOLICA-NORDEX N-60 Power curve [FOLDER].txt

19.062 EOLICA-NORDEX N-60 Technical description [FOLDER].txt

11.695 EOLICA-NORDEX N-60 Transport, roads, crane requirements [FOLDER].txt

2.213 EOLICA-NORDEX N-62 Annual energy yield [FOLDER].txt

4.474 EOLICA-NORDEX N-62 Power curve [FOLDER].txt

18.734 EOLICA-NORDEX N-62 Technical description [FOLDER].txt

10.681 EOLICA-NORDEX N-62 Transport, roads, crane requirements [FOLDER].txt

2.470 EOLICA-Nordic 1000, foundation data, preliminary [SHEET].txt

17.339 EOLICA-Opportunities and Prospects for Wind Power Generation.txt

28.368 EOLICA-Permitting of Wind Energy Facilities - A HANDBOOK.txt

37.597 EOLICA-Political prices or political quantities - A comparison of renewable energy support systems.txt

31.639 EOLICA-POWER GRID INTERCONNECTION - WIND - ASIA.txt

59.139 EOLICA-REPORT CONCERNING - PLANNING IN SPAIN.txt

5.309 EOLICA-Scope of the System Reliability Impact Study for the Canastota Wind Power Project.txt

480.088 EOLICA-STUDYING WIND ENERGY-BIRD INTERACTIONS - A GUIDANCE DOCUMENT METRICS AND METHODS FOR DETERMINING OR MONITORING POTENTIAL IMPACTS ON BIRDS AT EXISTING AND PROPOSED WIND ENERGY SITE.txt

22.150 EOLICA-Technical specification - Nordic 1000-54 - Manual.txt

16.007 EOLICA-The Economics of Wind Energy.txt

10.594 EOLICA-When the wind of change starts blowing there are who build walls and there are others building windmills - Hermann Scheer.txt

20.016 EOLICA-WHySE Wind-Hydrogen Supply of Electricity Markets - Technology - Economic.txt

169.368 EOLICA-Wind Energy - Powering Economic - Development for Colorado.txt

8.401 EOLICA-Wind Energy Applications Training Symposium - 2002.txt

63.470 EOLICA-Wind Farm Monitoring - DOE-NREL.txt

4.735 EOLICA-Wind Farm Monitoring - Folder.txt

1.761 EOLICA-Wind measurement.txt
 126.236 EOLICA-WIND POWER TRANSMISSION CASE STUDIES PROJECT.txt
 75.063 EOLICA-Wind Resource Maps of Southern New England - PROJECT REPORT.txt
 269.579 EOLICA-WindForce12 - A BLUE PRINT TO ACHIEVE 12Percent OF THE WORLD'S ELECTRICITY FROM WIND POWER BY 2020.txt
 32.525 EOLICA-WWEC 2003 - The World Wind Energy Conference - Folder.txt
 39.326 EQUIPELETRICO-Improving Circuit Breaker Maintenance Management Tasks by Applying Mobile Agent Software Technology.txt
 39.393 EQUIPELETRICO-MAWS - Automatic Weather Stations [FOLDER].txt
 13.389 EQUIPELETRICO-mico-mico3 Electric Energy & Power meter [FOLDER].txt
 11.640 EQUIPELETRICO-SIEMENS - Solar module SP75 - FOLDER.txt
 11.620 EQUIPELETRICO-SYMPHONIE - Internet-Enabled Data Loggers - FOLDER.txt
 39.239 FOTOVOLTAICA-Energy Conversion Devices, Inc. - 5th Annual Electric Utilities Industry Conference.txt
 16.497 GD-Capstone MicroTurbine - 30 KW - LF - ENEDIS - FOLDER.txt
 12.647 GD-Capstone MicroTurbine Standard Warranty Terms of Coverage for Standard Equipment & Optional Systems - MANUAL.txt
 16.185 GD-Installation-Application Issues - Breakout Session - Engineering Analysis - Feasibility Studies.txt
 755.749 GD-INTEGRATED RESOURCE PLAN 2003 - Assuring a bright future for our customers.txt
 15.044 GD-Micro and Mini Turbine Technology - West Coast Energy Management Congress.txt
 10.957 GD-MICRO-POWER - Big solutions in Small Packages.txt
 11.399 GD-Microturbine Technology & Business Developments.txt
 12.875 GD-Microturbines - Activities within the Office of Distributed Energy Resources.txt
 62.253 GD-Modeling the Feasibility of Using Fuel Cells and Hydrogen Internal Combustion Engines in Remote Renewable Energy Systems - NREL.txt
 38.590 GD-POWER PARK WHITE PAPER.txt
 82.763 GD-Review of Combined Heat and Power Technologies.txt
 27.980 GD-SCE MICROTURBINE GENERATOR TESTING PROGRAM.txt
 44.853 GD-Solar Hydrogen Production by Electrolysis.txt
 11.843 GD-TELEDYNE TITANTM HM GENERATOR SERIES - Hydrogen-Oxygen Gas Systems.txt
 4.648 GD-THE PHOENIX PROJECT - SHIFTING FROM OIL TO HYDROGEN WITH WARTIME SPEED - Conclusions.txt
 106.819 GD-US Department of Energy - Strategic Plan for Distributed Energy Resources.txt
 18.506 GD-Workshop - The Potential for BCHP.txt
 191.929 GD-WORKSHOP ON INTERCONNECTING DISTRIBUTED GENERATION - DOE.txt
 40.497 GERACAO-FERC Electric Tariff - GENERATOR INTERCONNECTION PROCEDURE - ATTACHMENT K.txt
 177.184 GERACAO-Generation Interconnection Guidelines For The Dairyland Power Cooperative (DPC) Transmission System.txt
 194.951 GERACAO-STANDARD GENERATOR INTERCONNECTION PROCEDURES - ATTACHMENT 1.txt
 49.506 GERENCIAMIENTO-Project Management Team - Water Resources Management Decision Support System - Meeting Minutes.txt

36.386 GIS-Integration of remote sensing data and geographic information system technology for emergency managers and their applications at the Pacific Disaster Center.txt

34.856 GIS-NREL - Geographic Information Systems in Support of Wind Energy Activities at NREL.txt

19.679 HIDROGENIO-STUART ENERGY - DELIVERY SUCCES - THE EVOLUTION OF ENERGY [FOLDER].txt

14.035 HIDROGENIO-STUART ENERGY - HYDROGEN POWER SOLUTIONS - POWERING A NEW GENERATION [FOLDER].txt

89.197 ILUMINACAO-VISION 2020 - THE LIGHTING TECHNOLOGY ROADMAP - 1 20-YEAR INDUSTRY PLAN FOR LIGHTING TECHNOLOGY.txt

189.652 INFRAESTRUTURA-CALIFORNIA ENERGY COMMISSION - Electricity Infrastructure Assessment.txt

252.877 INVESTIMENTO-LONG TERM SIGNALS FOR INVESTMENT IN TRANSMISSION.txt

36.980 MANUTENCAO-Agent-Oriented Approach toWork Order Management for Circuit Breaker Maintenance.txt

48.865 MANUTENCAO-AUTOMATED CIRCUIT BREAKER MONITORING AND ANALYSIS.txt

84.262 MANUTENCAO-Creating Owner's Competitive Advantage Through Contractual Services - GE Power Systems.txt

28.181 MANUTENCAO-INFORMATION DIAGNOSTIC SYSTEM FOR HYDROPOER.txt

178.037 MANUTENCAO-INFORMATION MONITORING AND DIAGNOSTIC SYSTEM PROTOTYPE.txt

16.462 MEDICAO-8300ION Advanced Socket-Mount Meter - FOLDER.txt

64.903 MEDICAO-8500 - 8400ION Advanced Socket-Mount Meter - FOLDER.txt

168.803 MEIOAMBIENTE-INQUIRY ETCr1920 - Trading in greenhouse gas emissions.txt

18.151 MEIOAMBIENTE-Practical Application of the Kyoto Mechanisms - Opportunities and Issues.txt

93.117 MERCADO-ALTERNATE STRUCTURES FOR PUBLIC POWER SYSTEMS.txt

94.841 MERCADO-Business process model for invoicing in the downstream electricity power market.txt

257.677 MERCADO-NETWORK PRICING INFORMATION PACKAGE.txt

182.194 MERCADO-Regional Electricity Markets Interconnections - Phase I - Identification of Issues for the Development of Regional Power Markets in South America.txt

104.767 MUDANCACLIM-A NET-ZERO FOSSIL FUEL USE HOME - CASE STUDY.txt

4.772 MUDANCACLIM-Association of University Leaders for a Sustainable Future - The Talloires Declaration - 10 Point Action Plan.txt

109.189 MUDANCACLIM-BC HYDRO - Submission of B.C. Hydro's 1997 Comprehensive Greenhouse Gas Action Plan.txt

91.130 MUDANCACLIM-ERU-PT - Emission Reduction Unit Procurement Tender - Terms of Reference.txt

116.261 MUDANCACLIM-Handbook on Methodologies for Technology Needs Assessments.txt

23.967 MUDANCACLIM-Hot air vs CDM - Pelangi paper - Limiting supply to make Kyoto work without the United States.txt

34.393 MUDANCACLIM-SGR - Cleaner technologies - a positive choice.txt

36.474 MUDANCACLIM-Verification of the Kyoto Protocol - A Fundamental Requirement Key Issues for the Sixth Conference of the Parties to the Convention on Climate Change.txt

43.041 NANOTECNOLOGIA-Biologically Related Aspects of Nanoparticles, Nanostructured Materials, and Nanodevices - Chapter 7.txt

33.870 NUCLEAR-DEVELOPMENT OF A HYBRID INTELLIGENT SYSTEM FOR ON-LINE MONITORING OF NUCLEAR POWER PLANT OPERATIONS.txt

24.561 NUCLEAR-Nuclear Power- NMAC (Nuclear Maintenance Application Center).txt

109.960 PCH-Low Head-Low Power Hydropower Resource Assessment of the North Atlantic and Middle Atlantic Hydrologic Regions.txt

477.936 PED-Activities Report from the Power Systems Engineering Research Center.txt

177.269 PED-Fuzzy Logic and Linear Programming Find Optimal Solutions for Meteorological Problems.txt

66.483 PED-Fuzzy Systems Applications to Power Systems - Chapter IV.txt

93.458 PED-INVENTORY OF BACKUP GENERATORS IN THE STATE OF CALIFORNIA.txt

101.060 PED-MODERNIZING THE NATIONAL ELECTRIC POWER GRID.txt

76.783 PED-NEW APPROACH TO POWER SYSTEM TOPOLOGY VERIFICATION.txt

304.399 PED-Preliminary Research and Development Roadmap for Protecting and Assuring the Energy Infrastructure.txt

22.547 PED-STATE OF CALIFORNIA - THE RESOURCES AGENCY - CALIFORNIA ENERGY COMMISSION - STAGE TWO PIER PROGRAM AREAS.txt

116.789 PED-STRATEGIC RESEARCH AND DEVELOPMENT - 2001 - 2002 HIGHLIGHTS.txt

19.401 PED-Study Committee 35 Annual Report.txt

12.481 PED-T&D Technology Research Stem - Proposed Research Areas for 2002 - 2005.txt

719.644 PED-Universal Interconnection Technology Workshop Proceedings.txt

7.183 PED-UWIG Program Activities - Member Initiatives & Networking Incentives & Opportunities for NW Public Power Sector.txt

7.230 PLANEJAMENTO-To Support a Provision in Final Version of the Nations Comprehensive Energy Plan to Provide a Hydropower and Wind Energy Plan.txt

16.302 POLITICA-A PUBLICATION OF THE BONNEVILLE POWER ADMINISTRATION - KEEP CURRENT - April 2001 Working together to keep the lights on and costs down.txt

30.358 POLITICA-Price Behavior (and Forecasting and Elasticities) in the Pulp and Paper Industry.txt

185.251 POLITICARENOV-action - DEVELOPING ENERGY ACTION PLAN.txt

115.607 POLITICARENOV-ACTION PLAN CALIFORNIA FOR DEMAND RESPONSE.txt

55.540 QUALIDENERG-A Novel Software Implementation Concept for Power Quality Study.txt

34.508 QUALIDENERG-Advanced Software Developments for Automated Power Quality Assessment Using DFR Data.txt

33.161 QUALIDENERG-AUTOMATED ANALYSIS OF POWER QUALITY DISTURBANCES.txt

12.254 QUALIDENERG-Electricity Management Grounded.txt

30.526 QUALIDENERG-ION - Application note - Preventing Downtime.txt

7.168 QUALIDENERG-Power Harmonics.txt

41.491 QUALIDENERG-THE NEXT GENERATION SYSTEM FOR AUTOMATED DFR FILE CLASSIFICATION.txt

52.513 QUALIDENERG-Use of Intelligent Techniques in the Power Quality Assessment Applications.txt

22.408 RENOVAVEIS-ATTRIBUTES OF RETs PROJECTS - Small Hydro Project.txt

248.017 RENOVAVEIS-CALIFORNIA ENERGY COMMISSION - PUBLIC INTEREST ENERGY - RENEWABLE POWERS THE FUTURE.txt

130.913 RENOVAVEIS-Clean Energy Blueprint - A Smarter National Energy Policy for Today and the Future.txt

11.844 RENOVAVEIS-Colorado Wind & Distributed Energy - Renewables for Rural Prosperity.txt

56.819 RENOVAVEIS-current research in Victorian universities - energy-alternative energy.txt

26.635 RENOVAVEIS-Draft-Statute of the International Renewable Energy Agency (IRENA).txt

8.790 RENOVAVEIS-Ediel-EAN project for the downstream electrical power market.txt

59.222 RENOVAVEIS-ENER - IURE PROJECT PHASE III - ANALYSIS OF THE LEGISLATION REGARDING RENEWABLE ENERGY SOURCES IN THE E.U. MEMBER STATES.txt

409.329 RENOVAVEIS-Factors Relevant to Utility Integration of Intermittent Renewable Technologies - NREL.txt

51.933 RENOVAVEIS-InfoPoint Energy Efficiency Issue 1-03.txt

47.138 RENOVAVEIS-Memorandum for the Establishment of an International Renewable Energy Agency (IRENA) by EUROSOLAR, the European Association for Renewable Energy.txt

10.315 RENOVAVEIS-PARLIAMENTARIANS FOR GLOBAL ACTION - OVONIC - Technology Responses for Sustainable Development [Apresenta#Eo].txt

148.533 RENOVAVEIS-Policy Options for Promoting Wind Energy Development in California - A Report to the Governor and State Legislature.txt

13.355 RENOVAVEIS-Presentation for the Global Solar and Wind Resource Assessment Project Wind Resource Assessment and Mapping at the U.S. National Renewable Energy Laboratory - NREL.txt

18.233 RENOVAVEIS-Proform - overview - A Tool for Assessment of Renewable Energy and Energy Efficiency Projects.txt

49.288 RENOVAVEIS-Renewable Energy - Overcoming Intermittency.txt

494.688 RENOVAVEIS-U.S Department of Energy - National Renewable Energy Laboratory - 2002 Research Review.txt

241.879 RENOVAVEIS-UNDERSTANDING NON-RESIDENTIAL DEMAND FOR GREEN POWER.txt

27.665 SENSORES-TOSSIM System Description - SIMULATOR.txt

69.829 SENSORES-Wireless Sensor Networks for Habitat Monitoring.txt

93.873 SETORELETRICO-What International Investors Look For When Investing In Developing Countries - Results from a Survey of International Investors in the Power Sector.txt

173.061 SISTEMPOT-AUTOMATED OPERATING PROCEDURES FOR TRANSFER LIMITS.txt

25.893 SISTEMPOT-SIMPOW - Power system simulation software [FOLDER].txt

94.124 SISTEMPOT-The California Energy Commission's Reporting, Forecasting & Data Collection Responsibilities.txt

3.821 SISTEMPOT-Views on the Northeast Asia Power Grid Interconnection.txt

5.973 SOLAR-Intersolar 2002 - 80 % of exhibition space booked Germany's largest solar fair almost booked out.txt

57.044 TERMELETRICA-DESIGN CONSIDERATIONS FOR HEATED GAS FUEL - GE Power Systems.txt

354.328 TERMELETRICA-Economic and Technical Considerations for Combined-Cycle Performance-Enhancement Options - GE Power Systems.txt

63.674 TERMELETRICA-GAS TURBINE PERFORMANCE CHARACTERISTICS - GE Power Systems.txt

58.095 TERMELETRICA-GATECYCLE PERFORMANCE ANALYSIS OF THE LM2500 GAS TURBINE UTILIZING LOW HEATING VALUE FUELS.txt

80.539 TERMELETRICA-GE Aeroderivative Gas Turbines - Design and Operating Features - GE Power Systems.txt

3.712 TERMELETRICA-GE Launches Outage Advantage for turbines and generators worldwide.txt

85.482 TERMELETRICA-GE- SPEEDTRONIC MARK V GAS TURBINE CONTROL SYSTEM.txt

132.480 TERMELETRICA-Heavy-Duty Gas Turbine Operating and Maintenance Considerations - GE Power Systems.txt

48.679 TI-MARKET POTENTIAL ANALYSIS - A METHODOLOGY FOR ESTIMATING THE MARKET POTENTIAL FOR COMPUTERS AND OTHER INFORMATION TECHNOLOGIES.txt

15.302 TI-STANFORD UNIVERSITY-TECHNOLOGY STRATEGY SUPPORT.txt

51.973 TI-vigil - Enforcing Security in Ubiquitous environments.txt

111.275 TI-Vigil - Providing Turs for Enhanced Security in Pervasive systems.txt

42.489 TI-Web Caching in Pervasive Computing World.txt

186.874 TRANSMISSAO-AEP Transmission Planning Project Proposal.txt

3.573 TRANSMISSAO-BCHYDRO - PROCESS GUIDE - WHOLESAL TRANSMISSION SERVICES.txt

- 133.144 TRANSMISSAO-California Electric Rule 21 - Supplemental Review Guideline.txt
- 104.670 TRANSMISSAO-Crystal 500kV Transmission Line Project - Environmental Assessment - Chapter 2 - Proposed Action.txt
- 179.183 TRANSMISSAO-TRANSMISSION LINE SAFETY AND NUISANCE.txt
- 94.416 TURBINA-SIMULATION METHODS USED TO ANALYZE THE PERFORMANCE OF THE GE PG6541B GAS TURBINE UTILIZING LOW HEATING VALUE FUELS.txt
- 24.653 TURBINA-TURBINE - Key factors affecting maintenance planning [IDEA MAPPING] - GE [veirifcar].txt
- 51.662 UHE-SAT HydroPower - Integrated Process Control Technology for Hydroelectric Power Stations.txt