**Extração de caracteristicas de palavras em textos na língua portuguesa em maquinas distintas através de algoritmos** ​***stemmer***

**Emerson Barbosa da Cunha** emerson.barbosa20@gmail.com

**Instituto Federal de Educação, Ciência e Tecnologia da Paraíba(IFPB) *Campus****​* **Campina Grande**

**1. Introdução**

Diante da enorme necessidade atual de se obter informações de documentos e textos de determinada língua para diversas finalidades, se vê uma necessidade de algorítimos eficientes para extração de radicais das palavras a fim de se poder ter um estudo em cima dos mesmos, esses algoritmos também podem ser chamados de ​***stemmer****​*.

Este trabalho visa comparar alguns dos algoritmos de radicalização que podem ser utilizados na língua portuguesa, uma vez que os mesmo impactam diretamente na qualidade do processo de mineração.

Os algoritmos que serão analizados estão entre os mais utilizados na língua portuguesa são eles: PORTER [PORTER 2007], RSLP [ORENGO, HUYCK 2001] e SAVOY [SAVOY 2006], utilizando suas implementações para a língua portuguesa.

Para a obtenção dos textos a serem analizados será implementada uma arquitetura de ambiente distribuído vizando obter arquivos texto em diversas maquinas onde as mesmas irão processar os dados e por fim armazenar em um local comum onde se poderá ser feita uma analise profunda dos dados e utilizados para outros fins.

**2. Mineração de Texto**

Mineração de Textos é um conjunto de técnicas com o objetivo de extrair de textos não estruturados ou semi­estruturados, informação inovadora e até então desconhecida. O processo de MT pode ser dividido minimamente em três etapas: preparação dos dados textuais, processamento do texto e avaliação dos resultados [GOMES 2006]

**3. Algoritmos** ​***stemmer***

**3.1 Porter** ​***Stemmer***

Originalmente escrito como um método de extração de radicais para a língua inglesa com a vantagem de ter uma construção razoavelmente simples e resultados com tempos

computacionais interessantes [PORTER 1980], posteriormente o método foi adaptado para a língua portuguesa [PORTER 2007].

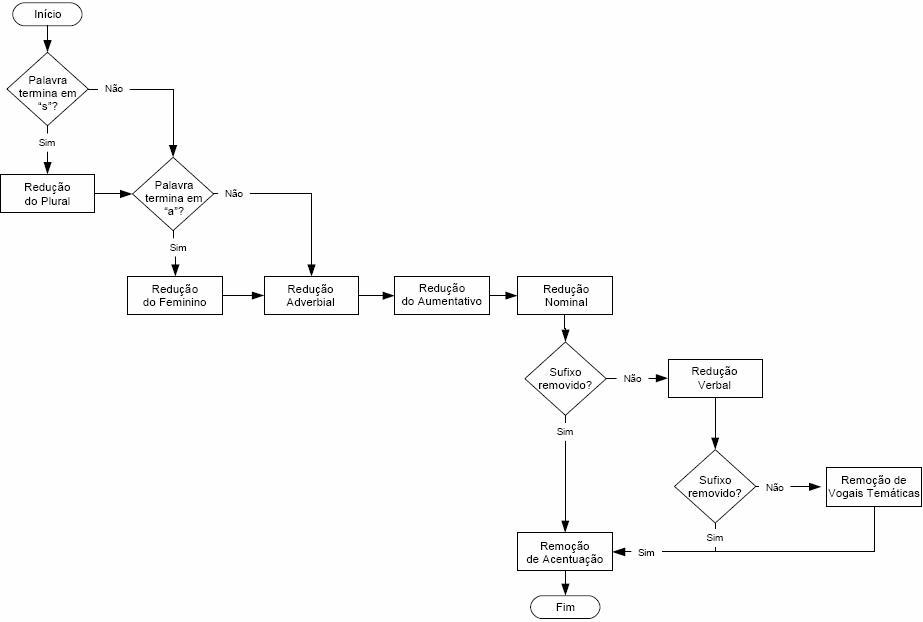
O algorítimo consiste em 5 passos:

* + Passo 1: remoção de sufixos comuns;
  + Passo 2: remoção de sufixos verbais, caso a palavra não tenha sido alterada no passo 1;
  + Passo 3: remoção do sufixo “i”, se precedido de “c” e se a palavra foi alterada pelos passos 1 ou 2;
  + Passo 4: remoção de sufixos residuais (os, a, i, o, á, í, ó) se nenhum dos passos anteriores alterou a palavra;
  + Passo 5: remoção dos sufixos “e”, “é” e “ê”, tratamento do cedilha e das sílabas “gue”, “gué” e “guê”.

1. **RSLP**

Proposto por [ORENGO, HUYCK 2001], foi projetado especialmente para a língua portuguesa, com objetivo de ser simultaneamente simples e eficaz melhorando a precisão de consultas em grupos de documentos. O diferencial deste algoritmo é que além de analisar diversas regras específicas do idioma português, ainda conta com um dicionário de exceções [XAVIER, SILVA, GOMES 2013].

**Figura 1. Diagrama de Fluxo do algoritmo RSLP [ORENGO, HUYCK 2001]**



**3.3 Savoy** ​***Stemmer***

Originalmente escrito para a língua francesa[SAVOY 1993], posteriormente adaptado para o português, alemão e húngaro [SAVOY 2006], é o mais simples entre os algoritmos que foram descritos e foi desenvolvido com o intuito de ser leve.

Ele é dividido nas seguintes etapas:

* + Passo 1: remoção do plural e do sufixo adverbial "mente";
  + Passo 2: transformação de gênero feminino para masculino;
  + Passo 3: retirar a vogal temática da palavra mantendo um mínimo de 3 caracteres;
  + Passo 4: remoção da acentuação da palavra.

1. **Comparação**

O método escolhido para a comparaçao dos algoritimos foi o metodo de Paice [PAICE 1994], este método busca medir os algoritmos de radicalização, avaliando dois erros comuns neste processo: ​***overstemming*** *​*e​***understemming*** *​*[PAICE 1994]:

* + ***Overstemming****​*:palavras que são convertidas para o mesmo radical mas tem conceitosdiferentes.
  + ***Understemming****​*:são palavras de mesmo conceito que são reduzidas a radicaisdiferentes.

1. **Implementação**

O objetivo deste trabalho é utilizar um conceito de ambiente distribuídos onde o programa seria disparado em varias maquinas e iria varrer as mesmas em busca de arquivos do tipo texto, onde a partir dai seria os textos seriam processados.

Os arquivos texto seriam processados pelos algorítimos stemmer e os radicais armazenados em um banco de dados, onde posteriormente poderiam serem utilizados para vários outros fins entre eles a comparação entre os algoritmos para se ter uma noção de qual deles teve um percentual mais baixo de erros.

**6. Referencias**

GOMES, G. R. Integração de Repositórios de Sistemas de Bibliotecas Digitais e de Sistemas de Aprendizagem. Tese ( Doutorado em Informática ) – Pontifícia Universidade Católica, Rio de Janeiro, 2006.

ORENGO, V. M.; HUYCK, C. A Stemming Algorithm for the Portuguese Language. 8th International Symposium On String Processing And Information Retrieval (spire), Laguna de San Raphael, Chile, 2001.

PAICE, C. D. An evaluation method for stemming algorithms. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer­Verlag New York, 1994.

PORTER, M. F. An Algorithm for Suffix Stripping, 1980.

PORTER, M. F. Portuguese stemming algorithm. Disponível em <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>. Acesso em 03/05/2015, 2007.

SAVOY, J. Stemming of French words based on grammatical categories. Journal of the American Society for Information Science, 1993.

SAVOY, J. Light stemming approaches for the French, Portuguese, German and Hungarian languages. Proceedings ACM.­SAC, 2006.

XAVIER, Bruno Missi; SILVA, Alcione Dias da; GOMES, Geórgia Regina Rodrigues. Analise Comparativa de Algoritmos de Redução de Radicais e Sua Importância Para a Mineração de Texto, Rio de Janeiro, 2013.