UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP

Instituto de Ciências Exatas e Biológicas – ICEB

Departamento de Computação - DECOM

BCC449 - Recuperação de Informação na Web 2017-1 Prof. Álvaro R. Pereira Jr.

Trabalho Prático

Parte 3

Nome: André Ribeiro de Brito

Matricula: 11.2.4985

A terceira parte do trabalho foi relatar o experimento, como parâmetros usados ou considerações relevantes. Tendo como informações necessárias para compreensão do que foi feito e do resultado.

O segue abaixo o pseduo código:

Leitura do arquivo, sendo:

File file = new File(C:\Users\andre\ri\Arq.json");

String arquivo = FileUtils.readFileToString(file, "utf-8");

Passo 2: Após a leitura, o arquivo ira conter um conjunto enorme de string, assim a única maneira de fazer a o casamento é usando o algoritmo Boyer Moore como citado abaixo. Porém o java tem o método chamado “split” que tem como função quebrar uma string em sub string, sendo há um ganho de processamento.

Passo 3: tendo agora um conjunto de sub string, transformo todo conjunto em letras minúscula, usando o método : toLowerCase.

Passo 4: o java tem um método chamado “matcher”, que tem como objetivo combinar a palavra com o texto. Sendo assim, foi utilizado nesse trabalho o método citado acima para verificar se o documento é ou não contrato.

O conjunto de palavras utilizadas como padrão foram: “contrato.*?” ,”locação.*?”, “.*?aluguel comercial.*?”, etc.

A utilização do “.*?” foi como base para verificar se aquele documento é um contrato de verdade, sendo que um contrato é “padrão + uma string” e um modelo de contrato é basicamente “padrão + _____”.

Treino	Teste	% Contrato	%Não Contrato	% Falso-Positivo	%Falso Negativo	Precisão	Revocação	F1
20	30	14	16	7	13	0.666	0.518	0.582

Obs.: A princípio foi utilizado o algoritmo Boyer-Moore, pois já tinha utilizado para fazer um trabalho semelhante da disciplina Estrutura de Dados. Esse algoritmo é uma heurística de casamento faz com que, ao mover o padrão para a direita, a janela em questão casa com o pedaço do texto anteriormente casado. Assim nesse trabalho passava um A=[string] e o B=[texto], ou seja , A seria a palavra que desejo e B o texto.

Segue o link do material:

http://www.decom.ufop.br/guilherme/BCC203/geral/ed2_casamento-cadeias.pdf

https://docs.oracle.com/javase/7/docs/api/java/util/regex/Matcher.html

Para calcular a precisão, revocação e f1, foi baseada na fórmula abaixo: