Universidade Federal de Uberlândia

Sistemas de Informação

Organização e Recuperação da Informação

Aluna: Beatriz Ribeiro Borges

Matrícula: 12021BSI231

# TP03 – Relatório

## Exercício 1

1) Execute o código passo a passo usando o Google Colab e crie um pequeno relatório (PDF) explicando o que foi feito. Explique a importância de pré-processamento, o conceito de stopwords e as diferenças entre stemming e lemmatization.

Uma vez baixado o corpus ‘machado’ é possível analisar os textos que serão tratados. Na linha 11 é baixado o texto “Dom Casmurro”, ou seja, sem nenhum tratamento ainda. Em seguida é feito o primeiro tratamento selecionando apenas as letras: significa remover a pontuação, números e outros caracteres. Depois o texto já é convertido para letras minúsculas.

A linha 13 baixa uma lista de stopwords em português e em seguida na linha 14 cria-se seu vocabulário usando set. Usando o set de stopwords do NLTK, são removidas as palavras que não possuem valor semântico para a busca (exemplo: artigos e preposições). Na linha 18 o laço verifica se a palavra está no set de stopwords, caso não esteja é colocado em uma nova lista. Por fim no pré-processamento, começa o trabalho de reduzir as palavras aos seus radicais. Nas linhas 20 e 21 trabalha-se com stemming para chegar a sua base.

Os gráficos e prints seguintes no código mostram a importância do pré-processamento, uma vez que sem ele pode haver resultados consideravelmente diferentes: os gráficos mostram diferentes palavras mais frequentes com e sem o tratamento. Com o processamento, normalizamos as palavras para que o programa consiga compará-las facilmente e encontrar melhores resultados. Aqui também é importante ressaltar a diferença de stemming e lemmatization. As duas formas trabalham para chegar ao radical da palavra, mas o stemming "corta" as palavras usando a raiz como base (usado no programa) e lemmatization reduz as palavras a forma verdadeira da raiz. É possível analisar as diferenças usando o gerúndio ‘andando’ como exemplo: com stemming ficaria ‘anda’ e com lemmatization ‘andar’. Por ser uma função mais complexa, lemmatization pode demorar um pouco para ser executado, caso tenha um texto muito grande. O nltk possui duas funções para stemming: Porter (PorterStemmer) – usado no código - e Lancaster (LancasterStemmer). Há, para o lemmatization, o lemmatizador (WordNetLemmatizer).