UNIVERSIDADE FEDERAL DE UBERLÂNDIA
SISTEMAS DE INFORMAÇÃO

Roberto Costa Tupinambá
Yan Gustavo Pegyn Silva
Guilherme Francisco Silva Vidigal

RELATÓRIO GERAÇÃO DE UM SISTEMA DE ORI

MONTE CARMELO
2020

Sumário

Objetivos

Dados vários arquivos de documento em formato PDF contento textos devemos realizar uma lematização, remover as stopwords e apresentar os resultados na forma estrutural de dados de índice invertido.

Descrição do problema

Em resumo, o problema consiste em lermos os documentos contendo os textos a ser tematizados e um arquivo das stopwords, tais palavras que se encontram no documento de stopwords não passaram pelo processo de lematização e serão, portanto, removidas do conjunto de documentos, sendo assim, o restante das palavras serão aplicadas alterações na flexão de gênero, ou seja, passando do feminino para o masculino e a remoção do gênero gramatical.

Estruturas de dados e limitações

Para a solução do problema nos utilizamos de estruturas de dados, sendo estas, listas e dicionários, e também simulamos um padrão de lista em arquivos de texto.

A solução implementada para o problema tem como sua principal limitação o não reconhecimento de inúmeras exceções de regras de normalização de palavras e também não conhece grande parte das “stopwords” existentes.

Funções utilizadas

Utilizamos a biblioteca “re” (Regex) nativa do python 3.9 e a biblioteca fitz que faz parte do pacote PyMuPDF.

Definimos 7 funções novas e a função principal, sendo elas:

1. read: Efetua o processamento do texto.
2. une_stopwords: Une todas as listas de stopwords do objeto de resultados e remove as repetições.
3. une_palavras: Une todas as listas de palavras do objeto de resultados e remove as repetições.
4. calcula_indice_invertido: Cria uma estrutura de dicionário com os dados necessários para montar o índice invertido.

5. formata_indice_invertido: Transforma o dicionário no texto formatado do índice invertido.
6. formata_lista_documentos: Cria o texto da lista de documentos.
7. formata_stopwords: Cria o texto das stopwords encontradas.

Já na função principal (__main__), é definido os nomes dos documentos que se deseja efetuar a análise, executa as funções elaboradas para processar os dados, e assim que finalizado esses processos salva em arquivos de texto os resultados obtidos.