Atlântico Bootcamp

NLP

VISÃO GERAL

Uma empresa contratante deseja estabelecer termos de maior relevância em um documento específico. Neste caso, considere o histórico de exames, consultas e procedimentos realizados por um paciente. Um sistema deve ser desenvolvido para que o médico possa ter uma visão geral do histórico do paciente sem a necessidade de analisar documento por documento. Com base nesta importância, vamos desenvolver uma etapa deste sistema. Tokenizar um texto, realizar remoção de stopwords, aplicar o processo de lematização e fazer uma análise quantitativa e visual subjetiva deste.

OBJETIVOS

1. Carregar o conjunto de documentos em PDF e armazená-los em alguma estrutura de dados.
2. Realizar o pré-processamento destes (tokenização e remoção de stop words, deixar todos os caracteres minúsculos...).
3. Lematização com a Lib stanza
4. Implementar para determinar as seguintes informações dos resultados obtidos em 3 :

    4.1 Term Frequency (TF):

    $$TF = qtd\ de\ ocorrência\ do\ termo\ em\ um\ texto\ /\ quantidade\ total\ de\ palavras\ do\ texto$$

    4.2 Document Frequency (DF)

    $$DF = qtd\ de\ ocorrência\ do\ termo\ em\ um\ conjunto\ de\ documentos$$

    4.3 Inverse Document Frequency (IDF)

    $$IDF = log(qtd\ de\ documentos\ /\ (DF + 1))$$

    4.4 TF-IDF

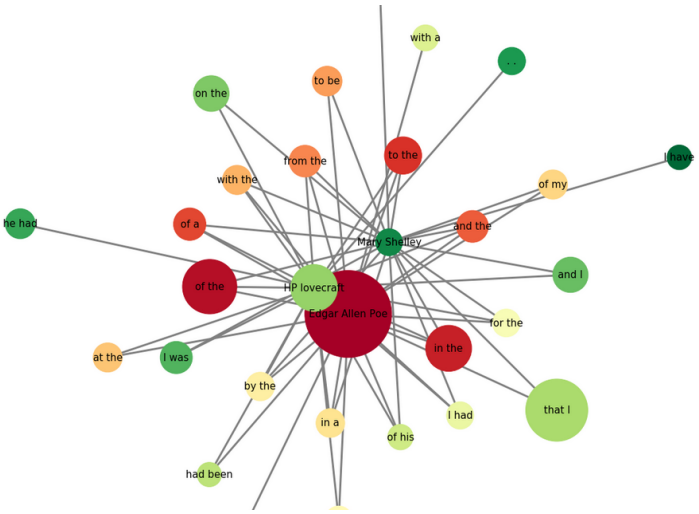
    $$TF - IDF = IDF * TF$$

    4.5 Lista de strings com proximidade até 2 dos 5 termos de maior TF-IDF. Essas strings devem ser acompanhadas de seu valor de TF. Exemplo: Suponha que a lista dos 5 termos de

maior TF-IDF é [casa, carro, comida, cachorro, gato]. Carro em um uma frase pode ter pneu e banco com as palavras mais próximas. Em outra parte do texto, carro pode ter volante e cinto, como as palavras mais próximas. Neste caso, para o termo carro, as strings [pneu,banco,volante,cinto] são as que devem ser armazenadas para análise.

5. Gerar um arquivo csv que possui todas as palavras de todos os documentos na primeira coluna, em que cada linha é um token. Para cada token, informe nas colunas vizinhas as informações determinadas no objetivo 4.1 até 4.4.

6. Gerar nuvem de palavras para análise visual tal como exemplo abaixo. Cada ponto central será um dos 5 termos de maior TF-IDF. As conexões são as palavras próximas obtidas em 4.5. O tamanho do círculo da palavra é baseado no TF dela. O maior círculo que conecta o termo central será normalizado para palavras de maior TF do conjunto.



Tópicos de Auxílio

Informações sobre as métricas utilizadas

https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089

Atividade determinação da nuvem de palavras

https://www.kaggle.com/arthurtok/ghastly-network-and-d3-js-force-directed-graphs

http://andrewtrick.com/stormlight_network.html