

# 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 deste. Neste caso, vamos comparar duas estratégias, se possível. A primeira utilizando a lib stanza e a segunda uma análise com base em acesso a um dicionário.

## 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. **Lematização manual com inspiração no trabalho descrito no link seguinte:**

<https://github.com/rikarudo/LemPORT> (Atividade desafio)

5. Implementar API para determinar as seguintes informações do resultados obtidos em 3 e/ou 4 :

5.1 Term Frequency (TF):

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

5.2 Document Frequency (DF)

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

5.3 Inverse Document Frequency (IDF)

---

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

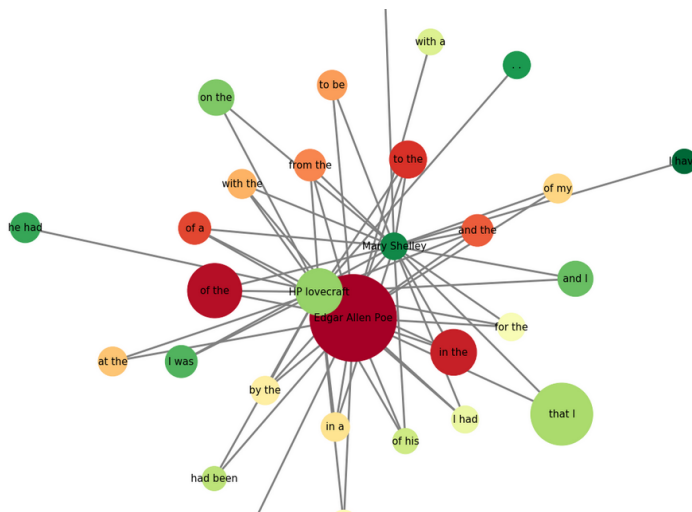
#### 5.4 TF-IDF

$$TF - IDF = IDF * TF$$

5.4 Lista de strings com proximidade até 2 dos 5 termos de maior TF-IDF. Essas strings devem ser acompanhadas de seu valor de TF.

6. 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 5.

7. 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 5.4. 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. (desafio)



---

## Tópicos de Auxílio

Se realizada a lematização manual, os resultados seguintes são duplicados e a comparação será realizada analisando a nuvem de palavras de cada implementação.

### Informações sobre as métricas utilizadas

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

### Atividade desafio de determinação da nuvem de palavras

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

[http://andrewtrick.com/stormlight\\_network.html](http://andrewtrick.com/stormlight_network.html)