**Universidade do Minho**
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

# Identificação de documentos similares/ relevantes para cancro do estômago

Perfil de Ciência de Dados
Aprendizagem Automática II
1º/4º Ano, 2º Semestre
Ano letivo 2020/2021

Orientadores : Rúben Rodrigues e Nuno Alves

*Autores:*
Hugo Faria, PG44415
João Soares, PG42836
Simão Gonçalves, PG42850

8 de junho de 2021

# Conteúdo

# 1 Introdução

Este projecto insere-se na unidade curricular de Aprendizagem Automática II, onde se pretende aperfeiçoar os conhecimentos obtidos na mesma, de forma a aplicá-los num contexto prático. Posto isto, foi-nos fornecida uma pipeline desenvolvida na framework do grupo BioSystems (BioTMpy) em que a mesma contém diversos documentos científicos sobre cancro do estômago. No entanto a maioria destes documentos não se encontra associado a uma sub-doença, sendo que, os dados que nos foram fornecidos encontram-se no formato JSON e foram-nos disponibilizados a partir da plataforma de conhecimento sobre cancro gástrico desenvolvido no grupo BioSystems. Assim, o objectivo que nos foi proposto com a realização deste trabalho trata-se de utilizar a pipeline que já se encontra desenvolvida na framework do grupo BioSystems (BioTMpy) e identificar a que sub-doença pertence cada documento não classificado. Desta forma, iremos implementar uma pipeline em python permita classificar documentos sobre cancro do estômago num conjunto de sub-doenças.Para isso iremos utilizar o LDA para calcular a similaridade de documentos para cancro de estômago recorrendo ao LDA e à fórmula de Jensen-Shannon, sendo que devemos integrar a pipeline desenvolvida na framework do grupo BioSystems (BioTMpy).

# 2 Plano de Resolução

Vista e percebida a principal a contextualização do nosso problema, começaremos agora a iniciar a resolução do que nos foi proposto. Assim, inicialmente, o grupo de trabalho definiu um conjuntos de passos estruturados para a resolução do problema. As tarefas definidas foram as seguintes:

- Criação do *Dataset*;
- Exploração dos Dados;
- Pré-Tratamento de Dados;
- Construção dos modelos;
- Obtenção e análise crítica de resultados;

Veremos mais á frente cada um dos cinco pontos referidos de forma mais detalhada e pormenorizada.

# 3 Criação do *Dataset*

O primeiro passo que realizamos foi o da criação do dataset, sendo que para isso realizamos diversas etapas.

- Recolher informação da API para um ficheiro json
- Converter o ficheiro json em objetos python
- Converter os objectos python para pandas
- Exportar para o formato .CSV

# 4 Pré-Tratamento de Dados

De modo a optimizarmos os nossos modelos e a obtermos melhores resultados, realizamos um Pré-Tratamento de Dados. É de frisar que este processo foi desenvolvido inicialmente tendo em conta o processo de pré-tratamento de texto geral para todos os problemas de NLP. No entanto foi completo após a exploração de dados realizada na fase seguinte.

- **Remover palavras com uma letra**

  Nos documentos utilizados, encontram-se presentes muitas palavras que contêm apenas uma letra e podem prejudicar o nosso objectivo. Posto isto, procedemos então á remoção destas palavras.

- **Remover números**

  Procedemos também á eliminação dos números presentes nos nossos documentos, já que estes não nos iriam ajudar para a identificação de uma sub-doença.

- **Vetorização de texto**

  Um modelo de machine learning não consegue interpretar o significado das palavras numa frase. Por isso temos de arranjar forma de converter para uma linguagem que ele consiga interpretar. Para isso vetorizamos as palavras existentes o utilizando o word_tokenizer. O que o processo de vetorização faz é converter as palavras para um vetor em que cada palavra é caracterizada pela frequência dessa mesma palavra numa dada frase.

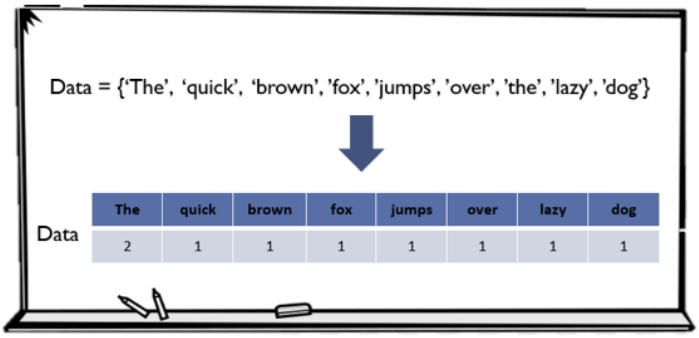


Figura 1: Exemplo de vetorização de texto

- **Utilização do isalpha**

  Os tokens foram guardados em alpha caso sejam apenas cosntituídos por letras do alfabeto. A função isalpha() retorna True, se todos os carateres forem letras do alfabeto. Por outro lado, retorna False, se o token contém um ou mais carateres que não pertencem ao alfabeto.

- **Remover stop-words**

  As stop words são palavras que em geral aparecem muitas vezes e não trazem benefício à tarefa que estamos a desenvolver sendo que por isto, retiramos as palavras mais frequentes da língua inglesa como "i"ou "me". Assim, reduzimos em muito as palavras das publicações sendo que por exemplo na publicação 1, o número de palavras foi reduzido, de 544 para 324.

- **Lemmatization**

  A lemmatization é o processo de agrupar as diferentes formas flexionadas de uma palavra para que possam ser analisadas como um único item. Assim, além de reduzirmos a dimensão das palavras conseguimos reduzir também a complexidade do treino e consequentemente aumentar o desempenho da tarefa de identificação de sub doenças. Podemos ver um exemplo em [**?**].

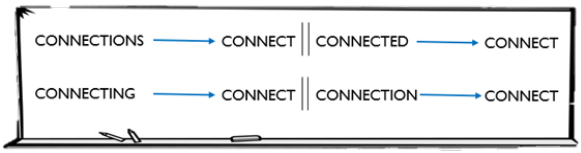


Figura 2: Correlação dos Atributos do Dataset de "Dados históricos da Bitcoin"

# 5 Exploração dos Dados

Na fase de exploração de dados consideramos fundamental avaliar vários aspectos do nosso conjunto de dados.

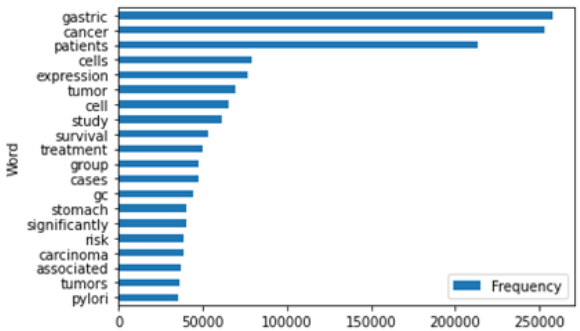


Figura 3: TOP 20 – Palavras Mais Usadas

Através desta figura podemos observar que existem palavras muito utilizadas que não nos ajudam a conseguir identificar a que sub-doença pertence um determinado documento pois são palavras comuns às várias sub-doenças.

Agora veremos a distribuição da frequência das palavras e os quartis obtidos.

| | |
|---|---|
| 0.00 | 1.0 |
| 0.25 | 1.0 |
| 0.50 | 1.0 |
| 0.75 | 3.0 |
| 1.00 | 257759.0 |

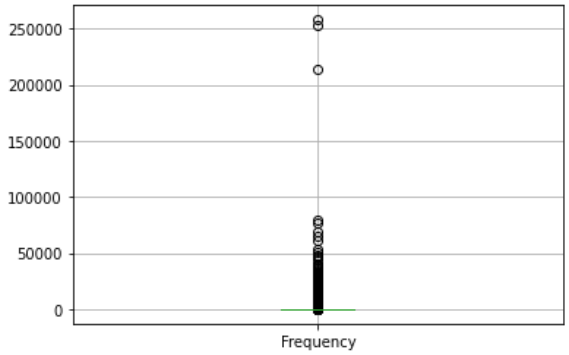


Figura 4: Distribuição da frequência das palavras e Quartis obtidos

Como podemos perceber 75%a das palavras aparecem apenas uma vez, o que é um valor extremamente reduzido comparativamente com o valor máximo que é 257668. Por isso poderá ser importante para os modelos que vamos construir a seguir utilizar esta análise exploratória.

# 6 Construção dos modelos

De modo a resolvermos o problema de classificação de documentos decidimos implementar uma abordagem de modelação de tópicos com o objectivo de associar as publicações de cada sub-doença a um dado tópico. Para esse efeito recorremos ao LDA

**LATENT DIRICHLET ALLOCATION (LDA)**

O LDA pressupõe que os documentos são compostos de palavras que ajudam a determinar os tópicos e os mapeia para uma lista, atribuindo cada palavra do documento a diferentes tópicos. Assim são calculadas a probabilidade de pertença de um documento a cada tópico e determinadas as palavras que representam um determinado tópico. É de salientar que uma das entradas importantes para o LDA é o número de tópicos esperados nos documentos de forma a criar o mesmo número de clusters que o número de tópicos.

# 7 Obtenção e análise crítica de resultados

Depois de criado um modelo que serviu como base, criamos outros modelos adicionando novas palavras. Mesmo após várias tentativas para na criação de modelos, não

conseguimos criar modelos que permitissem classificar os tópicos correctamente. Para além dos modelos que apresentamos, criamos outros como alterações na frequência de palavras removidas mas sem sucesso.

No entanto o nosso objectivo após obtermos um bom modelo capaz de distinguir os tópicos pelas várias sub-doenças seria aplicar a distancia de cada documento a cada tópico tornando essas distancias num vector de características associadas a uma label que iria representar a sub-doença de um documento. Com isto seria possível treinar um classificador supervisionado como o SVM para identificar a sub-doença de um documento usando a distancia a cada um dos tópicos do LDA

# 8 Conclusão

Com este trabalho podemos introduzir ao NLP, explorando e tratando um dataset textual de publicações de sub-doenças. No entanto não conseguimos cumprir o nosso objetivo principal que era criar um bom modelo LDA capaz de associar um documento a um tópico.