

Quick analysis of drug information

António Augusto Fernandes Simões Pereira

Nº 21136 – Regime Pós-laboral

Orientação

Prof. Joaquim Gonçalves

Prof. Patrícia Leite

Ano letivo 2022/2023

Licenciatura em Engenharia de Sistemas Informáticos

Escola Superior de Tecnologia

Instituto Politécnico do Cávado e do Ave

**Identificação do Aluno**

António Augusto Fernandes Simões Pereira

Aluno número 21136, regime pós-laboral

Licenciatura em Engenharia de Sistemas Informáticos

**Orientação**

Prof. Joaquim Gonçalves

[Categoria Académica do orientador]

Prof. Patrícia Leite

[Categoria Académica do orientador]

ÍNDICE

[1. Introdução X](#_Toc130363353)

[1.1. Objetivos 1](#_Toc130363354)

[1.2. Contexto 1](#_Toc130363355)

[1.3. Estrutura do documento 1](#_Toc130363356)

[2. Inclusão de Figuras e Tabelas 3](#_Toc130363357)

[3. Outros títulos 5](#_Toc130363358)

[3.1. Título de nível 2 5](#_Toc130363359)

[3.1.1. Título de nível 3 5](#_Toc130363360)

[3.1.1.1. Título de nível 4 5](#_Toc130363361)

[3.1.1.1.1. Título de nível 5 5](#_Toc130363362)

[4. Impressão 7](#_Toc130363363)

ÍNDICE DE FIGURAS

Siglas e Acrónimos

**NLP** – Natural Language Processing (Processamento de linguagem natural)

**DDS** – Drug-Drug Similarity

**LSTM** – Long Short-Term Memory

**HA-RCNN** - Hierarchical Attentive Recurrent Convolutional Neural Network

**DGNN** – Directed Graph Neural Network

**SRL** - Semantic Role Labelling

**TF-IDF** - Term Frequency-Inverse Document Frequency

# Introdução

O processamento de linguagem natural (Natural Language Processing) é uma das áreas da inteligência artificial que visa a interpretação e produção de linguagem natural (fala e escrita humana), por parte de computadores. Uma das funções do processamento de linguagem natural é a similaridade de texto que envolve medir o grau de semelhança entre dois textos. A similaridade de texto é extremamente importante para aplicações como verificação de plágio ou identificação de documentos duplicados, mas também pode ser aplicado a programas de reclamações ou de informações de medicamentos.

Neste projeto serão exploradas as bases da similaridade de texto e as suas principais técnicas e ferramentas utilizadas. Serão estudados diferentes algoritmos de processamento de linguagem natural e como estes podem ser utilizados para medir a similaridade entre textos de variados temas e explicação das abordagens para a avaliação desses algoritmos.

## Objetivos

O principal objetivo deste projeto é a assimilação dos vários conceitos da área da similaridade de texto.

Este projeto final de licenciatura orientado pelo professor Joaquim Gonçalves e pela professora Patrícia Leite

## Estrutura do documento

Em relação à estrutura do documento estes são os principais tópicos;

1. Analise do estado da arte (Artigos relacionados);
2. Análise e modelação do sistema, em que sejam levantados sistematicamente os requisitos, descritos diagramas de caso de uso e de atividade (que descrevam/formalizem o modelo de negócio).
3. Implementação, em que se descrevam as tecnologias escolhidas (e se justifiquem), e se refira detalhes sobre a implementação.
4. Análise de resultados e testes, seja uma análise/avaliação aos resultados obtidos, sejam testes de usabilidade ou unitários ao trabalho desenvolvido.
5. Conclusão

# Estado da arte

Com a crescente procura de aplicações que utilizem a similaridade de texto nos diversos setores, o investimento na pesquisa da mesma é extremamente importante para que novas tecnologias, que são necessárias no nosso quotidiano, sejam desenvolvidas e aprimoradas.

## Re-LSTM

O algoritmo Re-LSTM (Long Short-Term Memory) é um modelo de computação de similaridade de texto que utiliza uma rede neural recorrente (LSTM) para extrair características implícitas no texto. Este modelo utiliza uma abordagem de incorporação palavras ponderada através do algoritmo χ⊃2-C que avalia as diferentes palavras encontradas no texto através da sua categoria e frequência com que aparece no documento.

Outro método utilizado neste modelo é o TF-IDF (Term Frequency-Inverse Document Frequency) que é uma técnica utilizada no processamento de linguagem natural para medir a importância relativa de uma palavra presente em um ou vários documentos. Isto é possível calculando o peso de cada palavra com base na quantidade de vezes que aparece num documento específico ou vários documentos do mesmo tema. Se uma palavra aparece várias vezes num documento vai ter um peso maior pois é mais importante nesse documento específico.

## Drug-Drug Similarity

Drug-drug similarity (DDS) é um método que compara a semelhança entre dois ou mais medicamentos com base na sua estrutura química, propriedades físicas e químicas, alvos biológicos, mecanismos de ação e efeitos terapêuticos. Normalmente é utilizada em estudos de interações de medicamentos, onde a administração simultânea de dois ou mais medicamentos pode levar a efeitos colaterais indesejáveis ou diminuir a eficácia de um ou mais medicamentos.

Esta avaliação pode ajudar a identificar possíveis interações e fornecer informações para o desenvolvimento de novos medicamentos com perfis de segurança melhorados. Além disso, esta análise feita previamente pode ser usada para identificar medicamentos que possam ser reutilizados para tratar outras doenças ou para otimizar futuras receitas médicas com base no perfil do utente (historial de medicação e doenças associadas).

## A Short Text Similarity Calculation Method Combining Semantic and Headword Attention Mechanism

Neste artigo o modelo de rede neural usado é o Hierarchical Attentive Recurrent Convolutional Neural Network (HA-RCNN) que consiste na hierarquização das palavras em árvore através do seu “mecanismo de atenção” que se concentra nas partes importantes de um texto (safsadg, 2004).

Este mecanismo de atenção utilizado é chamado de headword, no qual é dada mais importância à palavra com mais peso de uma frase, melhorando a precisão da classificação de uma frase/palavra.

Apesar da reduzida dimensão da amostra utilizada no treino do modelo, os resultados indicam bom desempenho quando comparado com outros modelos que utilizam mecanismos de atenção. Os autores entendem que com uma amostra de maior dimensão os resultados podem, ainda, melhorar significativamente

Segundo os autores, este modelo teve um comportamento muito bom em relação a outros modelos com mecanismos de atenção, ainda que a dimensão da amostra utilizada no *dat*aset[[1]](#endnote-1)[[2]](#footnote-1)[[3]](#endnote-2) seja reduzida. o que os leva a acreditar que com uma amostra maior os resultados poderiam ser melhores também (ATEC utilizado para detetar fraudes discais e MSRP com o preço de carros indicado pelo fabricante e usado para tentar determinar o seu preço no futuro).

## Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information.

Neste artigo o modelo de rede neural utilizado é o Directed Graph Neural Network (DGNN) que trabalha com grafos de relação semântica (Semantic Role Labelling – SRL) para calcular a similaridade semantica do texto (Text Similarity).

Neste caso os autores fizeram dois testes um com SRL + DG (um grafo direcional convencional) e outro com SRL + SDG (um grafo em que todos os tipos de arestas são considerados como um único tipo).

Em termos de resultados os autores conseguiram observar que o segundo tipo de grafo aumenta o desempenho do algoritmo em relação a um grafo convencional.

## A fast text similarity measure for large document collections using multireference cosine and genetic algorithm.

Este artigo fala-nos sobre a importância da utilização da similaridade de texto para o aumento do desempenho dos motores de busca como por exemplo o Google.

Para que o desempenho de um motor de busca aumente é preciso remover todas as pesquisas com artigos/textos que sejam iguais ou similares para que diminua o número de resultados presentes no index resultando num tempo de pesquisa menor e menor probabilidade de encontrar informação repetida.

Os métodos utilizados neste tipo de texto similarity são chamados de Duplicate or Near-Duplicate – DND que, tal como o nome indica, são utilizados para encontrar artigos/textos idênticos para que não seja apresentada informação duplicada. Neste projeto foi utilizado o cosine texto similarity algorithm que cria, através da analise de várias partes de um artigo, um vetor term frequency-inverse document frequency (TF-IDF. Através deste vetor é possível saber se um outro texto que for analisado é igual ou parecido se o angulo dos vetores for parecido.

Bibliografia

# Bibliografia

Knuth, D. (1973). *The Art of Computer Programming.* Adison Wesley.

PennState University Libraries. (15 de Março de 2017). *APA Quick Citation Guide*. Obtido de PennState University Libraries Web Site: http://guides.libraries.psu.edu/apaquickguide/intext

safsadg. (2004). *asdfasdgf.*

1. [↑](#endnote-ref-1)
2. Um conjunto de dados para treino do modelo [↑](#footnote-ref-1)
3. [↑](#endnote-ref-2)