



Instituto Superior de Engenharia

Politécnico de Coimbra

DEPARTAMENTO DE / DEPARTMENT OF
INFORMÁTICA E SISTEMAS

Development of Learning Models in Biomedical Literature on a Collaborative Platform

Trabalho de Projeto para a obtenção do grau de Mestre em / to fulfill
the Master's degree in Informática

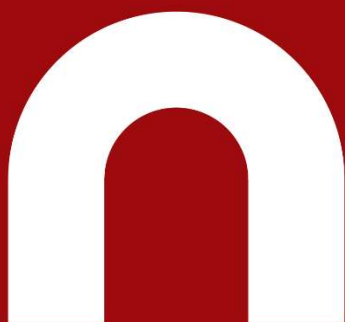
Especialização em Desenvolvimento de Software

Autor / Author

José Manuel Alves da Rocha e Barros

Orientador

Carlos Pereira



INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, October 2024

RESUMO

Ao longo das últimas décadas, o avanço científico realizado no campo de aprendizagem automática, em particular na área de processamento de linguagem natural, tem permitindo a modelos o processamento e consequente compreensão da linguagem humana. Este projeto envolve duas componentes, a primeira é dedicada à investigação e desenvolvimento de modelos com base em técnicas de aprendizagem profunda para a realização do processamento de linguagem natural com vista ao reconhecimento de entidades biomédicas, nomeadamente proteínas, genes e doenças. O conjunto de dados JNLPBA, BC2GM e NCBI Disease são referências para o reconhecimento de entidades nomeadas (REN) na área biomédica. O JNLPBA, permite a identificação de entidades do tipo ADN, tipo e linha de célula e proteínas, o BC2GM destina-se a identificar menções de genes, e o NCBI Disease contém anotações de nomes de doenças.

Foi desenvolvido um modelo NER para cada conjunto de dados utilizando a biblioteca SpaCy. O modelo treinado no conjunto de dados JNLPBA obteve uma pontuação F1 de 0,942, para o BC2GM uma pontuação F1 de 0,976 e para o NCBI Disease uma pontuação F1 de 0,986. Ao testar os modelos desenvolvidos num conjunto de quatro *abstracts* [1, 2, 3, 4], O modelo NER desenvolvido manteve pontuações perfeitas para os artigos 2 e 3, mas teve dificuldades com o artigo 4 no conjunto de dados BC2GM, obtendo uma pontuação F1 de 0. Em contraste, os modelos “en_ner_jnlpba_md” e SciBERT mostraram uma variabilidade significativa, particularmente com o reconhecimento de proteínas. O BioBERT foi coerente, mas variou no artigo 4. A fiabilidade e a eficácia do modelo desenvolvido fazem dele uma escolha superior para as tarefas NER.

A segunda componente deste projeto, é dedicada ao desenvolvimento de software visando a criação de uma aplicação web utilizando para isso a *framework* JHipster, resultando numa plataforma colaborativa que permite a utilização de modelos de aprendizagem, desenvolvidos pelos seus utilizadores que são investigadores da área da biologia computacional. Além da possibilidade de os utilizadores da plataforma adicionarem modelos, os modelos desenvolvidos na primeira componente deste trabalho, são disponibilizados na plataforma colaborativa. A plataforma descrita está em contínuo desenvolvimento, sendo que a descrição feita neste documento se refere à versão 1.0, e a mesma está disponível em www.geneticcollab.com [5].

Palavras-chave: Aprendizagem Automática, Processamento de Linguagem Natural, Aprendizagem Profunda, Reconhecimento do Nome de Entidade, Bioinformática, Biologia Computacional, Interação proteína-proteína, Desenvolvimento web

ABSTRACT

Over the last decades, scientific advances in the field of machine learning, especially in the area of natural language processing, have enabled models to process and understand human language. This project consists of two components, the first of which is dedicated to the research and development of models based on deep learning techniques to perform natural language processing to recognize biomedical entities, namely proteins, genes and diseases. The JNLPBA, BC2GM and NCBI Disease datasets are references for Named Entity Recognition (NER) in the biomedical field. The JNLPBA allows the identification of DNA type, cell type, cell line and protein entities, the BC2GM is intended to identify gene mentions, and the NCBI Disease contains annotations of disease names.

A NER model was developed for each dataset using the SpaCy library. The model trained on the JNLPBA dataset obtained an F1 score of 0.942, for BC2GM an F1 score of 0.976, and for NCBI Disease an F1 score of 0.986. When testing the developed models on a set of four abstracts [1, 2, 3, 4], the developed NER model maintained perfect scores for articles 2 and 3, but struggled with article 4 in the BC2GM dataset, achieving an F1 score of 0. In contrast, the "en_ner_jnlpba_md" and SciBERT models showed significant variability, particularly with protein recognition. BioBERT was consistent but varied in article 4. The developed model's reliability and effectiveness make it a superior choice for NER tasks.

The second component of this project is dedicated to software development aimed at creating a web application using the JHipster framework, resulting in a collaborative platform that allows the use of learning models developed by its users, who are researchers in the field of computational biology. In addition to the possibility for users of the platform to add models, the models developed in the first component of this work are made available on the collaborative platform.

The platform described is under continuous development, and the description given in this document refers to version 1.0, which is available at www.geneticscollab.com [5].

Keywords: Machine Learning, Natural Language Processing, Deep Learning, Named Entity Recognition, Bio-informatics, Biology, Protein-Protein Interaction, Web Development