Introducing Information Retrieval for Biomedical Informatics Students

Sanya B. Taneja, Richard D. Boyce, William T. Reynolds, Denis Newman-Griffis
University of Pittsburgh, Pittsburgh, PA, USA

Introduction

We developed a set of three activities introducing introductory BMI students to information retrieval with natural language processing (NLP), covering document representation strategies and language models from TF-IDF to BERT. These activities provide students with hands-on experience targeted towards common use cases and introduce fundamental components of NLP workflows for a wide variety of applications.

Learning goals

- Expose introductory BMI students to fundamental strategies for text representation and language models, geared towards information retrieval in biomedical contexts.
- Provide students with hands-on experience creating NLP workflows using pre-built tools.

Preprocessing
Inverted indexing
Information retrieval evaluation

→ Semantic similarity
Creating and analyzing embeddings

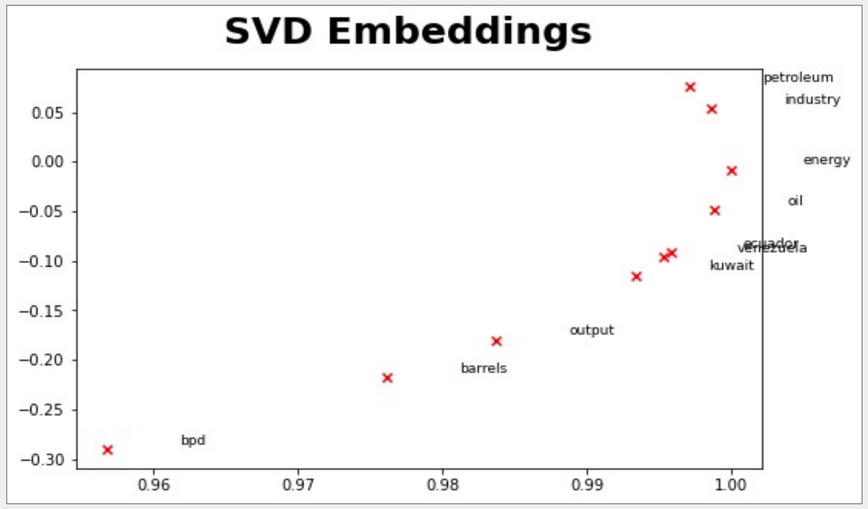
→ Advanced language models
Domain adaptation
Natural language inference

Notebook 1: Fundamentals of document analysis

- Basic preprocessing tasks in NLP workflows - tokenization, stemming, casing, and stop-word removal.
- Indexing techniques - inverted indexing and creation of a weighted document-term matrix using term frequency-inverse document frequency (TF-IDF).
- TREC evaluation measures including recall, precision, interpolated precision-recall average, and mean average precision.

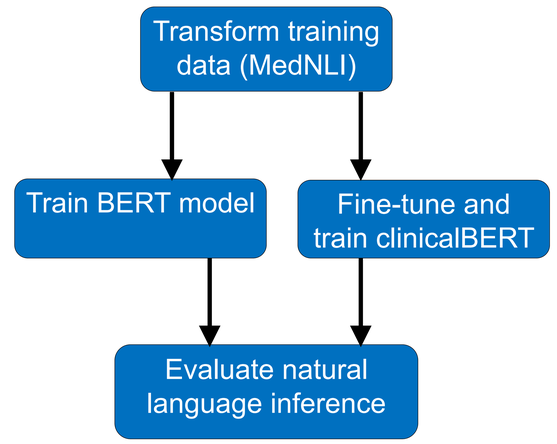
Notebook 2: Introduction to word embeddings

- Singular value decomposition (SVD)
- word2vec embeddings
- Visualization of embeddings



Notebook 3: Introduction to BERT and clinicalBERT

- Background on neural networks and design decisions in language models through YouTube tutorials.
- Named Entity Recognition and Medical Language Inference
- Compare performance of BERT and clinicalBERT using the Transformers library



Technical Skills and Libraries

- Python
- Jupyter notebook
- Natural Language Toolkit (NLTK)
- pytrec_eval
- Word2vec
- Gensim
- matplotlib
- Transformers by Hugging Face

Where to find the notebooks

https://github.com/dbmi-pitt/bioinf_teachingNLP

Challenges and Future Work

- Limited availability of benchmark datasets
- Limited annotations in publicly available clinical datasets (such as MIMIC)
- Ethical considerations of biomedical NLP regarding sensitive data
- More examples of domain specific language models and fine tuning
- Named Entity Recognition for biomedical texts using NLM-Chem

Discussion

- Activities to provide a modular workflow of components besides information retrieval; i.e., text preprocessing, indexing, execution, and evaluation.
- Reflect the perspective of the practical challenges that students face when working in biomedical NLP.
- Key to responsible use of NLP in BMI is determining what students need to learn and how to teach the information.

Department of Biomedical Informatics,
University of Pittsburgh

This work was supported in part by the National Library of Medicine of the National Institutes of Health under award number T15 LM007059.