User Guide for TMSA

Daniel Palacios

13 September 2022

Text Mining for Scientific Articles (TMSA)

TMSA is a small project written in python 3.10.7 that reads pdf and word document articles and performs generic Natural Language Processing (NLP) tasks. First it extracts the text data from a given word or pdf document, it uses a text preprocessing pipeline with well known NLP techniques like: Tokenization, removal of stopwords, removal of short words, Lemmatization, Stemming, removal of unnecessary characters. Note that there is a function where the user can manually add customized stop words for their specific article. After running a cleaning pipeline, TMSA proceeds to perform topic modeling using two different algorithms: Latent Dirichlet Allocation, and Non-negative Matrix Factorization. Topic modeling is the unsupervised process where an statistical model groups text in specific topics and extracts key concepts, ideas, and knowledge from the given text. The Term Frequency - Inverse Document Frequency (TF-IDF) scores are also displayed for the top words. Note that in this demo we consider single words, however, we can also explore bigrams or trigrams (2 or 3 word phrases) by manipulating the ngram range of the TfidfVectorizer() function from sklearn. TMSA also performs entity extraction and relation extraction with Spacy dependency relationships, which allows for the creation of static and interactive knowledge graphs using Bokeh. TMSA also creates a word cloud map to picture the statistical weight of dominant frequency word in the text.

Installation

Note that the software has been tested in several independent systems with Windows 10 or Windows 11. The testing was done in powershell or anaconda environments. For windows it is suggested to install python from the windows store.

Install the necessary dependencies with the following commands:

```
pip install pypdf2 nltk python-docx sklearn spacy
pip install tqdm networkx pandas numpy wordcloud bokeh
python -m spacy download en_core_web_sm
python -m nltk.downloader stopwords wordnet omw-1.4
```

Running the program

To run TMSA make sure you know the directory or path to the file you are trying to text mine, for simplicity we consider a file (docx or pdf) to be in the same directory.

To run we run the following command:

```
python TMSA.py
```

The command prompt will ask the user for an input. Here the exact name of the file has to be given. For this demo we consider the sample article provided: "aQTL_atlas.docx". A folder named Results will be generated in the current directory with all the outputs files from TMSA: Interactive graph html file, Top topics from LDA and NMF, a png image with the static relationships, TF-IDF histogram, and a png image with the word cloud content.