NATURAL LANGUAGE PROCESSING ON RESEARCH ARTICLES

Hung Cong Tran

January 27, 2023

Overview

1. Motivation and Problem identification
2. Data and Data Cleaning
3. Data Processing and Modelling
4. Future Improvements

Motivation and Problem identification

Nowadays most research articles are published online.

However, finding relevant articles may be difficult and time-consuming.

This fact creates a big demand for tagging/labeling each article by the related subject.

A good tagging/labeling system will contribute a lot to the search process and recommendation process not only in academia but also in industry and the economy.

Motivation and Problem identification

In this project, I build a model to recognize the subject of each article using the title and the abstract.

Data

The data contains the title and abstract of 20972 research papers from six subjects (Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, and Quantitative Finance).

Each paper is tagged as one, two, or three subjects.

Most papers (96%) in the data are tagged in four subjects Computer Science, Physics, Mathematics, and Statistics.

I decide to work on the classification problem in four subjects Computer Science, Physics, Mathematics, and Statistics

Data and Data Cleaning

I compute the title length, and the abstract length for each paper.

Then I create boxplots on these lengths from the whole papers and papers on each subject.

After that I remove the outliers from the data based on these boxplots.

Data and Data Cleaning

I write multiple functions to clear text that includes number removing, special character removal, punctuation removal, stop word removal, etc.

Then I apply these functions to clear texts in paper titles and abstracts.

I write a function to stemmized words and apply it to paper titles and abstracts.

Data and Data Cleaning

After performing data cleaning, the new data set contains 20243 papers with four more columns title (after cleaning), abstract (after cleaning), title length, and abstract length.

Data Processing

I perform the following data processing:

1. Try multiple tools to numerize texts in paper titles and paper abstracts that include Word2Vec, Doc2Vec, CountVectorizer, TfidfVectorizer, Doc2Vec, etc.
2. Train several quick models to measure the effectiveness of these text-processing tools.
3. Finally, choose the most effective one which is CountVectorizer.

Modeling

I fit the data with many traditional Machine Learning models that includes KNeighbors Classifier, Logistic regression, SVM (with polynomial kernel), Random Forrest, etc.

However, these models do not give me satisfying results (the best accuracy for all subject classifications is only 84% on the validation set.

Then I decide to build a Neutral Network with handy-pick architecture and parameter selections. The average accuracy reaches to 90% on the validation set.

Model performance on the test set

I measure the performance of the chosen model on the test sets and obtain the following results.

Report for the performance on the test set					
Subject	Accuracy	Precision	Recall	F1-score	AUC score
Computer Science	0.87	0.83	0.86	0.85	0.93
Physics	0.94	0.93	0.86	0.90	0.97
Mathematics	0.90	0.84	0.79	0.82	0.95
Statistics	0.88	0.77	0.74	0.75	0.94

Future Improvements

In the future, I plan to expand this project as follows:

1. Collect more research papers and create better NLP tools to process the text data.
2. Redesign the Neutral Network architect to create a better tagging model.
3. Develop the current model to a “bigger” one so that it can tag papers in more subjects.
4. Create a model that helps classify papers in my current research subject (Mathematics) based on many research areas such as algebra, geometry, analysis, topology, etc.

The End