NLP Project Proposal

Title: Document Similarity
Team Members: Siri Koduru, Lakshmi Yalamarthi

Paper Title: Aspect Based Document Similarity for Research Papers
Bibliographical Info: A publication at COLING 2020.
Authors: Malte Ostendroff, Terry Ruas, Till Blume, Bela Gipp, Georg Rehm.
Link: https://arxiv.org/abs/2010.06395

Background:
This paper describes a model to detect similarity between two documents and the aspect in which they are similar. Most of the traditional document similarity models predict the percentage at which two documents are similar or dissimilar, but this paper discusses the amount of similarity between a document pair based on each aspect of paper. Researchers rely merely on citations of papers to know the amount of similarity between two scientific publications, this is a binary classification. This paper enhances this idea into multi label and multi class classification. This classification is multi class because each paper has multiple sections, and a document can be cited in another document which makes it multi label classification.

Summary of Contributions:
The section title with citation acts as a label for the paper, the labels are preprocessed and normalized before labelling the datasets. Transformer based models like baseline LSTM, BERT models are used on multi label multi class pairwise documentation, to find text similarity between documents. For Aspect based classification, each document is divided into 11 label classes where each label indicates different sections of paper like Introduction, background, and conclusion, calculating F1-score, precision and recall on the individual labels of the paper shows in which aspects a document pair is similar. This paper performs negative sampling, by introducing a class name NONE which collects document pairs that are dissimilar when selected randomly. Several experiments with different transformer models are calculated on ACL Anthology and CORD-19 datasets. The experiments concluded that SciBERT has maximum prediction including the section of citation.

Limitations and Discussions:
This model worked well in document pairs with similar words in their introduction label and if the title and abstract of the paper had the same words. Results from some of the experiments have given more F1-score and Precision to 'None' class which shows dissimilarity in most of the documents in input dataset and number of positive labels are less and resulted in lower F1-score. The model has more accuracy and precision for multi labelled samples through the SciBERT model. Results from the second experiment shows

samples labelled as ‘related work’ with higher accuracy when compared to ‘introduction’ or ‘background’ labelled samples. Thus, showing reduced accuracy in the aspect-oriented results. Six other labels in ACL anthology datasets like Conclusion, Discussion have shown discrepancy in the results with lower accuracy. This model does not include images, scientific diagrams, graphs in the input datasets. Paper cites that correctness of the model is evaluated manually; a proper evaluation method will improve accuracy of the model.

Why this paper?

This paper is suggested by the professor during project proposal discussions. The major interesting aspect of this paper is negative sampling of input datasets, and four-fold classification between training and test datasets.

Wider Research Context:

The format of expressing a concept in a particular language differs from author to author, so finding similarities in the context of two discussions is a manual task, this can be achieved by algorithms like word to vector and cosine distance. This paper is built based on these models and developed using other transformation models.

Main goal of the project:

This project focuses on finding text similarities with text embedding and represents text into vector space. Our project uses similarity methods to find similarities in the granular level of the document. Our goal is to achieve document similarity as discussed in the previous paper and investigate the possibilities to improve accuracy with chosen input datasets. Success of the model helps in finding differences between original work and plagiarized ones.
Secondary goal of the project is to find similarity in the documents and pairing them based on their objectives discussed.

NLP tasks addressed:

We implement Cosine distance, Manhattan and Euclidean distance between word embeddings which shows similarities. For example, if input is two word documents our model will execute on the input pair, and outputs the amount of similarity between each paragraph of the document pair.

Input Datasets:

We have found a few datasets that might work for this project.
Letters to shareholders of Berkshire Hathway
This dataset contains 2000 documents.
We are planning to perform Normalization of data like removing punctuations and special characters, Tokenization, removal of stop words, stemming and lemmatization as part of preprocessing of data.
Apart from the above one we have also found out a 20 newsgroups data set from http://qwone.com/~jason/20Newsgroups/ which consists of around 18000 documents.

Methods planning to use:

We are planning to use word embedding transforms like TF-IDF, this algorithm will embed words into vector space. Distributed bag of words version of paragraph vector is another algorithm we are planning to implement, this algorithm receives paragraph id from document as input and it predicts occurrence words from document. We are planning to test the model with multiple methods like Doc2vec, TF-IDF and other transformers discussed in the reference paper like BERT, SciBERT to compare execution times for each method before picking the best performing one.

Baseline methods:

We are planning to split the data set as training, test and validation set and plan to use k-fold validation for testing the accuracy of the model.

Evaluation of Results:

We are planning to divide input datasets randomly into training, test and validation datasets with a ratio of 3:1(train and test sets ratio). We predict the accuracy of the model by calculating F1 score, Precision, and recall.