CA4022 Assignment 3 Midway Report

Project Outline

In this project we perform a comparative analysis of several word embedding approaches to effectively classify Yahoo's Answers Q&A into their respective topics. We will focus on the use of Natural Language Processing alongside supervised learning approaches.

Dataset

We will use the Yahoo Q&A dataset. This dataset has 4 variables; Class Index (1 - 10), Question Title (string), Question Content (string) and Best Answer (string). We will be predicting the Class Index variable which corresponds to one of the following topics:

- Society & Culture
- Science & Mathematics
- Health
- Education & Reference
- Computers & Internet
- Sports
- Business & Finance
- Entertainment & Music
- Family & Relationships
- Politics & Government

Each question and answer could have several topics, but in this dataset a question and answer's main topic is chosen as the Class Index. Each class contains 140,000 training samples and 6,000 test samples with 1.4 million observations in the training dataset and 60,000 observations in the test dataset.

Technology

For data preprocessing and modelling, we will use PySpark, an API for using Apache Spark in Python. PySpark will allow us to analyse our data and perform transformations as we would do in a distributed, big data architecture. The dataset we are using is large (~325MB, ~1.5m rows). Therefore we will use Hive for data size reduction, if necessary. Using Hive, we can produce a stratified sample of training data from each question topic in the dataset.

We will set up a connection between PySpark and Jupyter Lab to improve our workflow when running our code and carrying out experiments. Our code and documentation will be stored on a collaborative git repository located on the GitHub platform here.

Analytic Approach

There are several steps to our analytic approach. Firstly, we must preprocess the data. This involves manipulating or dropping data to prepare the dataset for further analysis such as stop word removal and tokenization so we can apply embeddings to the text fields in our data.. Secondly, we will perform feature engineering and feature extraction. We will figure out a way to join the question and answer fields into a "Docuemnt". After that, we will begin the comparison of word embedding methods such as Word2Vec/Doc2Vec, TF-IDF, BERT and GloVe. Finally, we will train a Linear Support Vector Machine and classify the data into their respective classes.

Responsibilities and Roles

We split up the tasks for this assignment as follows:

Research

This involves researching the different word embedding techniques and narrowing our analysis down to two. We will also learn the proper implementation of word embeddings and machine learning models in PySpark. We will both carry out research and discuss our findings.

Preprocessing

Diarmuid will do the data preprocessing step. Joseph will review this and we will both discuss the approaches used here to produce the best outcome.

Word embedding / Modelling

We will use one embedding approach each since we will focus on two out of the four possible word embeddings. We will compare our results before starting the writeup.

Report Writeup

Joseph will write up the final report, which will then be peer reviewed by Diarmuid.