DS 5230 Unsupervised Machine Learning
Project: Plagiarism Detection

Team Members:
Aishwarya Vantipuli
Spatika Krishnan
Nandini Jampala

Abstract:

Plagiarism is considered an academic dishonesty and breach of journalistic ethics. The problem of plagiarism is majorly faced in educational institutions. To protect one's academic integrity we are attempting to implement our project on Plagiarism Detection. The dataset we've chosen is publicly available on Corpus of Plagiarised Short Answers and is created by Paul Clough and Mark Stevenson. The corpus has been designed to represent varying degrees of plagiarism and what we envisage will be a useful addition to the set of resources available for the evaluation of plagiarism detection systems. The corpus consists of answers to the following five short questions on a variety of topics in the Computer Science field.

- What is inheritance in object-oriented programming?
- Explain the PageRank algorithm that is used by the Google search engine.
- Explain the Vector Space Model that is used for Information Retrieval.
- Explain Bayes Theorem from probability theory.
- What is dynamic programming?

Each question has answers of 19 students and 1 Wikipedia answer. Four levels of plagiarism are represented in the corpus: Near copy, Light revision, Heavy revision, Non-plagiarism. It is a multi-level classification problem.

The goal of this project is to find the list of students with plagiarised work for each question from the given answers of students as input using unsupervised machine learning methods such as K-means Clustering and FP-Growth Algorithm. Before applying a clustering algorithm to the dataset we create a word vector for each unique word using the GloVe method. So now when we implement the k-means algorithm we can know to which cluster each word vector belongs. Thus similar word vectors would belong to the same cluster. To find out how many sentences are plagiarised we use the FP growth algorithm. It is based on a prefix tree representation of the given database of transactions (called an FP-tree), which can save considerable amounts of memory for storing the transactions. In our project, each vector can be viewed as an item-set and each answer can be viewed as a transaction. Feeding these transactions into an FP-Growth Algorithm we can find which sentences are plagiarised by just looking at the transactions which in our case are student answers. As features obtained from Tf-Idf Vectorization are quite a lot, we are planning to use PCA for reducing dimensionality and for visualizing features in 2D space.