ARTIFICIAL INTELLIGENCE PROJECT

TOPIC: Plagiarism detector

Team Members:
21PD08 - G S PRETHIKA
21PD19 - KRITHIKA L

Abstract

In this project, we will be building a plagiarism detector that examines a text file and performs binary classification; labeling that file as either *plagiarized* or *not*, depending on how similar that text file is to a provided source text. Detecting plagiarism is an active area of research; the task is non-trivial and the differences between paraphrased answers and original work are often not so obvious.

One of the ways we might go about detecting plagiarism, is by computing similarity features that measure how similar a given text file is as compared to an original source text. We can develop as many features as we want and are required to define a couple as outlined in this paper. In this paper, researchers created features called containment and longest common subsequence.

We will be defining a few different similarity features to compare the two texts. Once we have extracted relevant features, we will explore different classification models and decide on a model that gives us the best performance on a test dataset.

Project Components:

- Notebook 1 : Data exploration
  - Load in the corpus of plagiarism text data
  - Explore the existing data features and the data distribution
- Notebook 2 : Feature Engineering
  - Clean and pre-process text data
  - Define features for comparing the similarity of an answer and a source text and extract similar features.
  - Select “good” features, by analyzing the correlations between different features.

    - Create train/test .csv files that hold the relevant features and class labels for train/test data points.

REFERENCE:

- https://s3.amazonaws.com/video.udacity-data.com/topher/2019/January/5c412841_developing-a-corpus-of-plagiarised-short-answers/developing-a-corpus-of-plagiarised-short-answers.pdf