Design Document for Plagiarism Checker

## Team:

* Chandrahas Aroori (f2016A7PS0100H)
* Naren Surampudi (f2016AAPS0206H)
* Aditya Srikanth (f2016A7PS0091H)

## Topic

Implementation of basic plagiarism checker using Python3.

## Introduction

**Plagiarism detection** is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet has made it easier to plagiarize the work of others. Most cases of plagiarism are found in academia, where documents are typically essays or reports. Our plagiarism checker checks the level of plagiarism between a given document and a corpus of documents.

## Libraries used

The python libraries used were:

1. Nltk: This library is generally used for Natural Language Processing, we have used it to tokenize the data and remove the stop words.
2. os: This library is generally used for System Commands, we have used it for get the current working directory.
3. pickle: This library is generally used for serializing objects in python, we used it to serialize dictionaries which stored our data.
4. Math: This library is generally used for mathematical functions in python, we used for the logarithmic function.

## Implementation

### Corpus Preprocessing

The tf-idf score was calculated was then calculated for each pair of (term, document) using the formula:

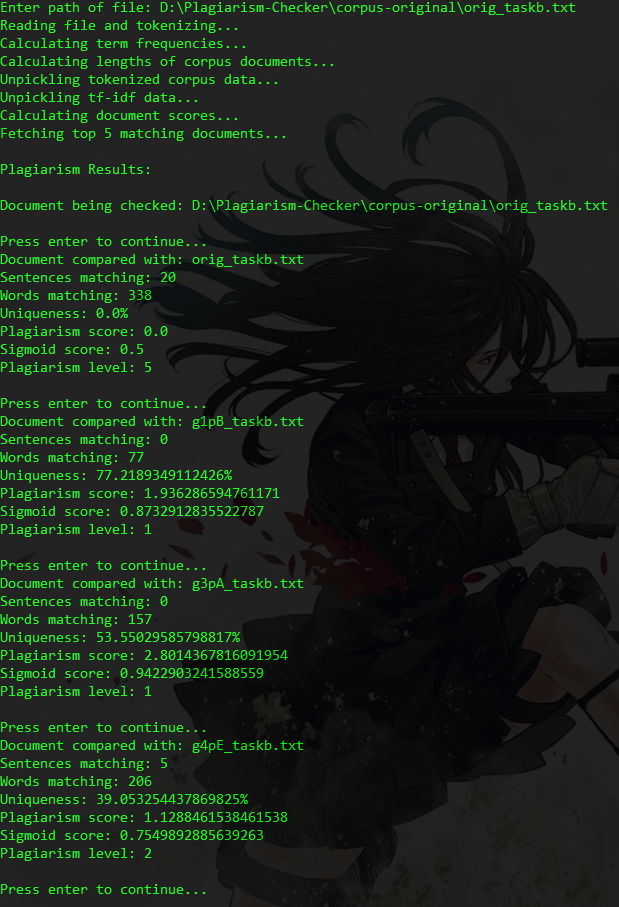
##### Pseudocode for Tfidf calculation:



### Query Processing and Vector Space Model

For the query processing we used nltk to tokenize the strings and then used regex for elimination of stop words. The term frequency is then calculated form this processed query vector. The vector space utilizes the tf-idf model to calculate the cosine score, which can be simply written as the product of the tf-idf score in the document and the corresponding term frequency in the query. This is then finally used to obtain the top documents which are then in turn passed through a check8ing algorithm to ch4eck for uniqueness and plagiarism level.

## Result



## References

1. Introduction to Information Retrieval; Christopher D. Manning, Prabhakar Raghavan, Hinrich Shutze; Cambridge

## Project Link

* https://github.com/nsurampu/Plagiarism-Checker