Indian Institute of Technology Bhubaneswar

Security & Forensics Lab II

Laboratory Experiment No. 5

(Held on 11/03/2022)

Submitted By:

Ayush Modi

21CS06005

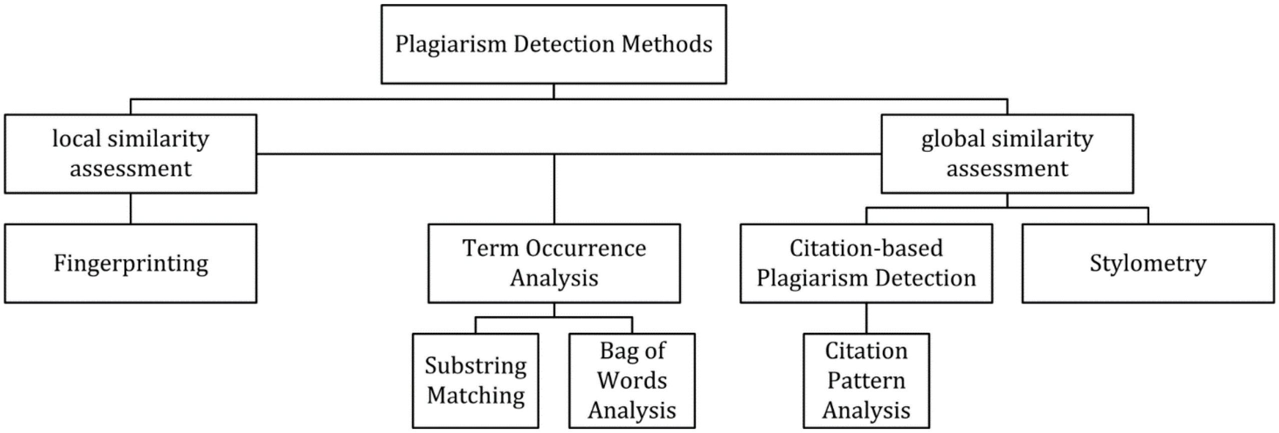
M. Tech, CSE (1st year)

Aim

To calculate the similarity index between two text files

Theory

Plagiarism detection or **content similarity** detection is the process of locating instances of plagiarism or copyright infringement within a work or document. The widespread use of computers and the advent of the Internet have made it easier to plagiarize the work of others. Text matching is the key issue in these applications. For matching between two texts, we need a representation of each text and a similarity measure between two texts. Vector Space Model is a typical model for text representation. For building better YSM, researchers have brought forward a variety of feature selection schemes, but for the text-matching task like job searching, the text features are very different from those in previous work. In these texts, a number of special names, such as job name, technical name, place name, and fixed words have a decisive role in the text matching. Most of these features are some types of multi-word expressions.


Plagiarism Detection Methods
local similarity assessment
global similarity assessment
Fingerprinting
Term Occurrence Analysis
Citation-based Plagiarism Detection
Stylometry
Substring Matching
Bag of Words Analysis
Citation Pattern Analysis

Here, we have adhered to Term Occurrence analysis.

Procedure

The following steps are performed to check the similarity between the files:

1. Reading both the base file and the file to be compared

3. Convert all characters to lower case in both the files

3. Replace punctuation with space in both the files

4. Get token by tokenizing the lines

5. Perform Stemming – To generate common form of each word
(Like – rock and rocks should be considered as the same word even though their tenses are different)

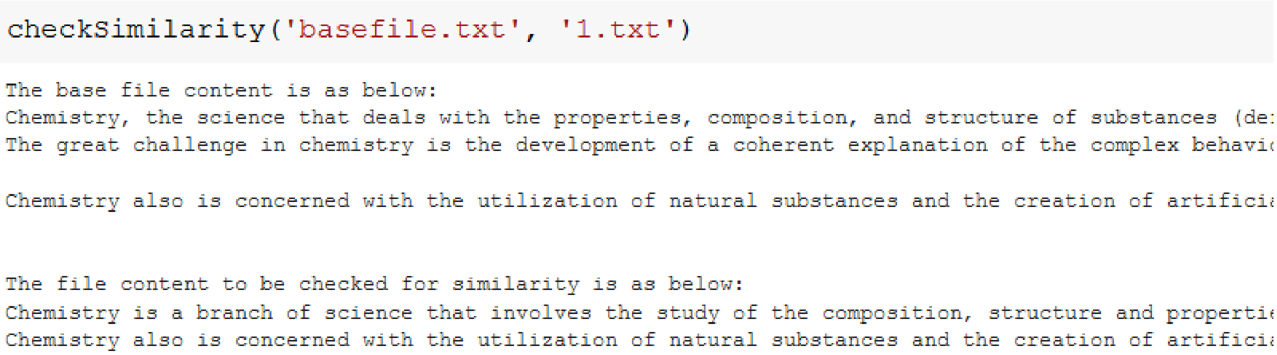
6. Now get the tfidf (**Term Frequency — Inverse Document Frequency**) for the input file and combine previous file contents.

7. Calculate the words that are similar between both files.

8. Calculate the plagiarism percentage.

Results

The following files were compared:



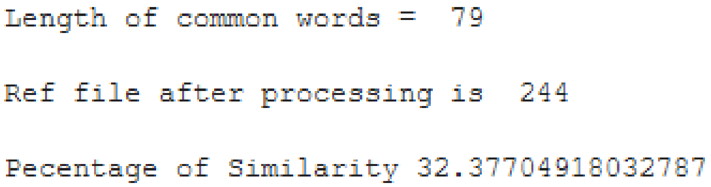
```
checkSimilarity('basefile.txt', '1.txt')
```

```
The base file content is as below:
Chemistry, the science that deals with the properties, composition, and structure of substances (de
The great challenge in chemistry is the development of a coherent explanation of the complex behavi

Chemistry also is concerned with the utilization of natural substances and the creation of artifici


The file content to be checked for similarity is as below:
Chemistry is a branch of science that involves the study of the composition, structure and propertie
Chemistry also is concerned with the utilization of natural substances and the creation of artifici
```

The following result was obtained:



```
Length of common words =  79

Ref file after processing is  244

Pecentage of Similarity 32.377049180327787
```

Conclusion

The text file was checked for plagiarism and similarity index with respect to the base file.

Code

Google drive link –
https://drive.google.com/drive/folders/1f-UiQtXgeIImVProqS0DIuDHS70YVmVJ?usp=sharing

Github link to the file –
https://github.com/21CS06005/Plagarism_Checker