Information Retrieval (CS F441) – Assignment 1

Plagiarism Checker

Team Members:

1. Surinderpal Singh Virk (2018A7PS0234H)
2. Arpit Adlakha (2018A7PS0250H)
3. Saksham Pandey (2018A7PS0259H)

Abstract:

Plagiarism is a term used when any anonymous person uses someone else's piece of work and consider that work of its own by not giving them proper credits either intentionally or not intentionally. It has become very often that students copy other students work in school and colleges.

Plagiarism is detected by Containment measure Technique. In this technique we use a dataset(a corpus of documents).

Steps followed for this developing this application:

Step 1: Packages required for our application to run are imported. Natural Language Toolkit(NLTK) is imported. NLTK is a set of libraries and program which provides stemmers, tokenizers, lemmatizers etc. Tkinter is used for making the GUI where the user inputs the text to be checked. Matplotlib is a plotting library, used to draw a graph for showing extent of plagiarism.

Step 2: After importing the packages we traverse through our corpus and input document and read it.

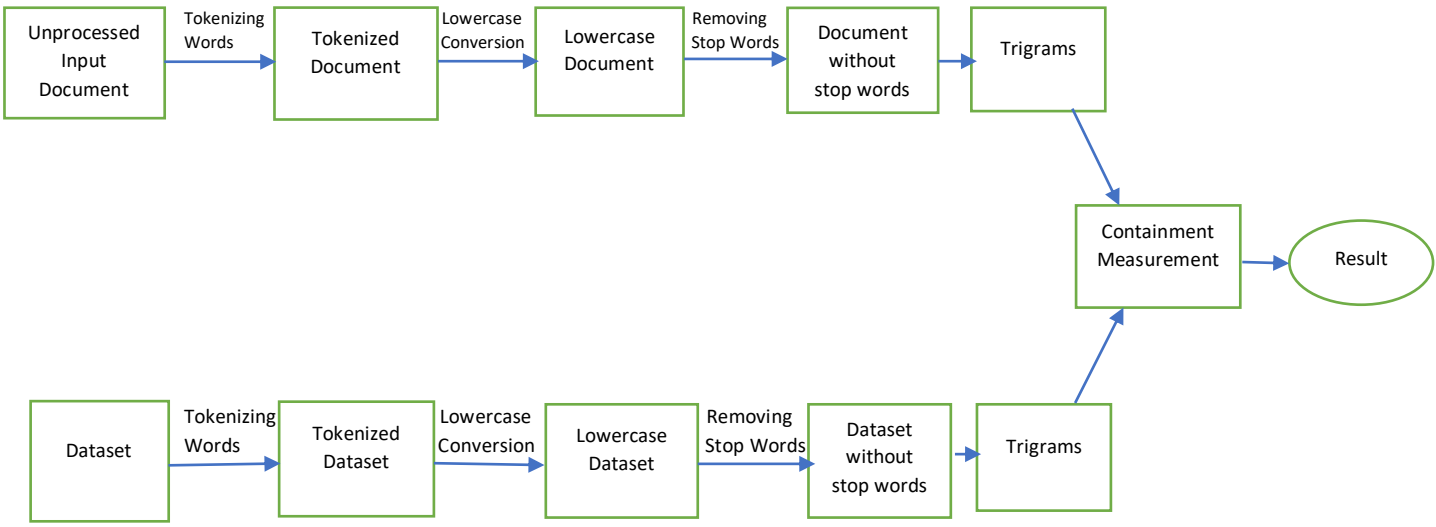
Step 3: After reading the corpus and input document we run a pre-processing on them and convert them to trigrams. Words are tokenized using word tokenizer. After this, we convert our words to lowercase words followed by removing the stop words. This whole comes under the pre-process step which is also shown in the flow diagram below.

Step 4: After pre-processing now we check the amount of similarity. The amount of similarity text between two texts is calculated with the help of Jaccard similarity coefficient or Containment measure C. But here we are using containment measure because it is a better measure for document pairs with varied lengths.

S(A): set of trigrams in our input document
S(B): set of trigrams in our dataset

The containment measure C(A,B) is given by,
C(A,B)=(Intersection of S(A) and S(B))/S(A)

Design and Architecture:



Result=C(A,B)