**Project A**

Calculate the similarity for each pair of text files listed in fintempsim.csv

Create a csv file and include the following four columns.

1. File 1 name (Provided in the csv file)

2. File 2 name (Provided in the csv file)

3. Similarity between File 1 and File 2

4. Similarity between File 1 and File 2 (remove stop words and the words listed at the end of the document)

**Project B**

Create a csv file and include the following nine items.

0. File name

1. How many words are there in each file?

2. What is the size of each file? (how many kb?)

3. Complexity of each file

4. What is the within-document lexical diversity of each file?

5. How many unique words are there in the file?

6. How readable is each file?

7. Count <1>“Supplemental” “Supplement” “Supplements” “supplemented”<2>”Annual”<3> “update”,”updates”,”updating”, “updated”

8. If there are less than 150 words but greater than 3 words in the file, print the first 30 words.

**def** clean(riskFactorString):  
 riskFactorString = riskFactorString.replace(**'Item 1B'**, **' '**)\  
 .replace (**'font'**,**' '**)\  
 .replace (**'Staff Comments'**,**' '**)\  
 .replace (**'Item 1A'**,**' '**)\  
 .replace (**'Table of Contents'**, **' '**)\  
 .replace(**'Unresolved Staff Comments'**, **' '**)\  
 .replace (**'Pagebreak'**, **' '**)\  
 .replace (**'END LOGICAL PAGE'**, **' '**)\  
 .replace(**'BEGIN LOGICAL PAGE'**, **' '**) \  
 .replace(**'also'**, **' '**) \  
 .replace(**'will'**, **' '**) \  
 .replace(**'may'**, **' '**) \  
 .replace(**'FOLIO'**, **' '**) \  
 .replace(**'SEQ'**, **' '**) \  
 .replace(**'Not applicable'**, **' '**).strip()