National University of Computer and Emerging Sciences, Lahore Campus

Course:	Information Retrieval	Course Code:	CS5040
Program:	MS (CS) / MS(DS)	Semester:	Spring 2022
		Total Marks:	100
Deadline:	11-September-2022		
Section:	MCS-1A		
Assessment	Assignment - 1		

Instruction/Notes:

1. Make sure to read the instructions carefully before attempting them.

2. You are allowed to discuss with TA for general advice.

3. You must submit your own work on Google Classroom.

4. **Plagiarism will be not tolerated**. Penalty will be decided later on.

5. You have to code your logic in attached python notebook

6. Submit your notebook renamed as RollNumber_01.ipynb (Make sure it is renamed with your roll number for example L12345_01.ipynb)

7. If you do not have anaconda installed, you can use google colab.

8. The code must be very clear, to-the-point, and presentable. We will not just be looking for correct answers rather highest grades will be awarded to write-ups that demonstrate a clear understanding of the material. Write your solutions as if you were explaining your answer to a colleague. Style matters and will be a factor in the grade.

9. **You can use NLTK python library to complete the tasks.**

Tasks:

1. Read the file and find total occurrences of word

1.1 Write a function that returns count of number of lines in txt file.

1.2 Write a function that returns count of number of words in txt file.

2. Write a function that returns a data frame with two columns having stop words in one col and their count in the other.

3. Now generate a .txt file named Rollnumber.txt having all content of the given file but excluding stop words.

4. Write a function that displays all words having length greater than 5.

5. Write a function that returns the count of all words which end at y.

6. Write a function that prints the count of lowercase words in the file.

7. Write a function ZICount() to count all occurrences of Z and I (including small cases z and i too).
   For example: Hi! I am Zenia. It was so good to see you!
   Output:
   I or i : 4
   Z or z : 1

8. Write a function to display the content of a file in descending order (based on the word count).
   Content: I am confused to start it but it won't be done until I give it a try
   Output: I it to a am be confused done give start try until wont.

9. Perform lemmatisation on the given file and generate a txt file named Rollnumber_lemmatized.txt having all content of the given file.

10. Perform stemming on the given file and generate a txt file named Rollnumber_stemming.txt having all content of the given file.

Good Luck!