

**CSE556: Natural Language Processing**  
**Assignment 01**  
**Deadline: Sep 19, 2020 11:59 P.M (Saturday)**

Max Marks: 100

**Instructions:**

- The assignment is to be attempted individually.
- Language allowed: Python
- You are allowed to use libraries such as NLTK for data preprocessing.
- For Plagiarism, institute policy will be followed. Refer: [Academic Dishonesty Policy](#)
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Output.pdf.
- Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
- Mention your sample outputs in the output.pdf.
- You are advised to prepare a well-documented code file.
- Submit code, readme, and output files in ZIP format with the following name:  
**A1\_<roll\_no>.zip**
- Use classroom discussion for any doubt.

Your task in this assignment is to write a python program that accepts as input any text file from the “rec.motorcycles” and “sci.med” folders in [20newsgroups](#) dataset and performs the following tasks.

- 1) Print the number of words and sentences contained in the file given as input.
- 2) Print the number of words starting with consonants and the number of words starting with vowels in the file given as input.
- 3) List all the email ids in the file given as input.
- 4) Print the sentences and number of sentences starting with a given word in an input file.
- 5) Print the sentences and number of sentences ending with a given word in an input file.
- 6) Given a word and a file as input, print the count of that word and sentences containing that word in the input file.
- 7) Given an input file, print the questions present, if any, in that file.
- 8) List the minutes and seconds mentioned in the date present in the file given as input. (For instance, for the date - Tue, 20 Apr 1993 17:51:16 GMT, the output should be 51 min, 16 sec)
- 9) List the abbreviations present in a file given as input.

**Note:**

- Your code will be checked for any file in the above-mentioned folders as input. So, don't try to fit your code too closely to a single file.

- The numerical values present in a file can also be treated in words. For instance, 100 may be searched as hundred.