NLP Project Round - 1

Goto http://www.gutenberg.org
Download two considerably large books in text format
Using Python do the following for both the books separately:

- Import the text, lets call it as T1 and T2 (books that you have downloaded)
- Perform simple text pre-processing steps and tokenize the text T1 and T2 — you may have to do the removal of running section / chapter names and so on. Explore the txt file you will come to know what I am saying.
- Analyze the frequency distribution of tokens in T1 and T2 separately
- Create a Word Cloud of T1 and T2 using the token that you have got
- Remove the stop words from T1 and T2 and then again create a word cloud - what's the difference it gives when you compare with word cloud got before the removal of stop words?
- Evaluate the relationship between the word length and frequency for both T1 and T2 — what's your result?
- Do PoS Tagging for both T1 and T2 using anyone of the four tag sets studied in the class and get the distribution of various tags

For all the above points, presenting your result with proper visualisation and inferences that you get from the visualisation is very important.

Prepare a complete report of the above proceedings with all the necessary details starting with data description, data preprocessing steps, data preparation, problem statement, plots, tables, figures and output with your inferences and submit it through Moodle. All codes also need to be submitted together with the report (through a link to Github).

In the first page of the report properly mention your Team name, Member names with their Roll Nos. And Submit this report as a PDF file in Moodle. Follow the file naming as <Team_Name>_Project_Round_1.pdf. Only PDF will be accepted and no other format will be accepted. Teams can be formed with three members in the team. There are 85 members in NLP Course as per MIS. So ***only one team*** can have four members.

Last date for the submission of this project is Nov 20, 2021 (11.59pm). No extension of the deadline is possible for any reasons. There will be heavy penalty for not submitting it on time.