Project Round 1

NLP Project Round - 1

Download a novel from Internet (in PDF Format) that contains at least 200 pages in it.

Convert the book from PDF to text format (you can use online web-based tools for this). Using Python do the following for both the books separately:

- Import the text, let's call it as T (book that you have downloaded and in TxT format)
- Perform simple text pre-processing steps — you may have to do the removal of running section / chapter names / remove the pictures / tables and so on. Explore T you will understand (you may have to see regular expressions to do this).
- Tokenise T and Remove the stop words from T
- Analyse the frequency distribution of tokens in T.
- Create a word cloud on the Tokens in T
- Do PoS Tagging for T using anyone of the four tag sets studied in the class and get the distribution of various tags.
- Take the largest chapter, say C, of the book and create a bi-gram probability table for this chapter. **For this you should not remove the stop words**.
- Take any chapter (other than C) and play the Shannon game – Play the fill-in the blanks game with the bi-gram probability learned from the previous step.
- See how much accurate it is by comparing with the original sentence.

Once you do the above steps you need to prepare the report with the following details

- First page of the report – your team name, team members name with roll numbers and the book name that you have taken for this project. Submit in the Moodle as a single PDF file.
- Prepare a complete report of the above proceedings with all the necessary details starting with data description, data pre-processing steps, data preparation, problem statement, plots, tables, figures, output with your inferences and conclusion. All codes also need to be submitted together with the report (through a link to Github).
- For all the above points, presenting your result with **proper visualisation** wherever necessary and **inferences** that you get from the **visualisation** is **very important.**
- When you submit the file in Moodle, follow the file naming as <Team_Name>_Project_Round_1.pdf. Only PDF will be accepted and no other formats will be accepted.
- Last date for the submission of this project is Oct 25, 2023 (11.59pm). No extension of the deadline is possible for any reasons. There will be **penalty** for **not** submitting it on time.