Natural Language Processing (CS5803): Assignment-1

Domain-Specific Web Scraping and Exploratory Text Analysis

Assigned on: 10th January, 2025
Deadline: 24th January, 2025

Problem Statement

In this assignment, you will create a document collection from a specific domain. Some examples are healthcare, education, entertainment, tourism, finance, electronics, etc. You can also choose any other domain of your choice. The collection should contain at least **300 documents**, but feel free to download more if you can. You may consider focusing on content in some non-English languages (this may pose some challenges with the existing tools that work well with English, but their performance in non-English languages is unknown).

Tasks

Task 1: Domain Selection and Web Scraping

- Choose a domain of interest.
- Identify websites related to the chosen domain.
- Crawl the websites to get the pages.
- Use web scraping tools such as `BeautifulSoup` to extract text content from the websites.
- Save the extracted data in a structured/semi-structured format. Make sure to include the URL and title along with the extracted content.

Task 2: Data Cleaning and Pre-processing

- Clean the raw text by removing HTML tags, special characters, numbers, and stop words.
- Tokenize the text into words and sentences.
- Apply stemming or lemmatization to reduce words to their root forms.

Task 3: Exploratory Data Analysis

- Perform word frequency analysis to identify the most common words in your dataset.
- Generate visualizations such as:
  - Bar charts of word frequencies
  - Distribution plots of sentence lengths
- Calculate additional statistics such as:
  - Average word length
  - Number of unique words
  - Lexical Diversity
  - Any other statistics or observations you may find interesting or useful.

Submission Guidelines

- Submit a zip file named `<YourRollNumber>.zip` that contains
  - Your Jupyter Notebook (.ipynb) file
  - The dataset you created (collection of documents) in a structured format (e.g., CSV, JSON, or text files).
- Ensure that your code is well-commented and easy to understand.

Additional Notes

- You may use any publicly accessible website for scraping, but ensure compliance with the website's terms of service.
- If you have any questions, contact the TA of the course at least 2 days before the deadline.

Happy Coding!