Text Retrieval & Search Engine (CP423B)

Assignment 3

Maximum Points: 100

Instructions-

- Here are the paraphrased instructions:
- You are required to work on this assignment in the same group as the previous one.
- Only Python language is permitted for this assignment.
- Plagiarism will be handled according to the institute's policy.
- Your submission should include a recorded video, README.pdf, and code files. It is important to provide clear and concise comments within the code.
- You are allowed to utilize libraries like Pandas and BeautifulSoup.
- In the README.pdf, please outline the methodology, preprocessing steps, and any assumptions made.
- Include a description of your outputs and any analysis conducted (if applicable) in the README.pdf.

Question 1- Introduction to Webscraping [80 points]

Your task involves working with the given URL, which pertains to historical population counts of various Canadian provinces as recorded on Wikipedia (https://en.wikipedia.org/wiki/List_of_Canadian_provinces_and_territories_by_historical_population). Your goal in this segment is to import HTML tables from the web and manipulate them using pandas.

Here's a breakdown of what you will do:

1. Retrieve the specified webpage as raw HTML using the requests library
2. Decode the HTML into a tree-structured Python object with the BeautifulSoup library
3. Utilize BeautifulSoup to identify and extract only the tables we're interested in
4. Merge the tables, sanitize the text, and transform them into a single Python dictionary
5. Construct a pandas dataframe out of this dictionary
6. Locate all h2 elements on the HTML page and display their text content
7. Generate a list of all the hyperlinks embedded within the tables
8. Download every webpage by traversing the links included in the list created in the previous step.

Note: [20 points]

- Record a video demonstrating your system.

Good luck!