Information Retrieval - Sessional 1

A. Mustafi

August 18, 2023

1. Read in the *Classic* corpus into a list, where each text should be an entry into the list. Roughly how long does it take to read in the entire corpus. You can answer in absolute time.

2. Repeat the process using the *bbc* dataset. However, files inside the folders need to be added to the same list. How long did it take to read the dataset?

3. Use the bbc dataset for this problem. How many raw tokens can you identify? A raw token is obtained by segmenting on a blank space. Any punctuations are retained. Multiple consecutive white spaces are to be collapsed into a single white space.

4. Produce a list of better quality tokens for the documents in the bbc dataset. The tokens are now defined as follows:

   > "A Token is a collection of alphabets and/or digits with a (optional) single instance of an apostrophe or a hyphen."

5. Create a vocabulary for the bbc dataset containing the tokens found in the last exercise.

6. Write a program that accepts a token and plots a bar graph showing the top 20 documents where the token is present. Top 20 here is defined as per frequency of occurrence.

7. Plot a histogram for the length of tokens in the vocabulary you have just created. What PDF does it seem to fit best?