Information Retrieval
Sessional - 3 (Set 2)

A. Mustafi

August 25, 2023

For the exercises in this set we shall work with the *bbc corpus*. In all the exercises a token shall imply a *set of characters containing only alphabets and digits with atmost one apostrophe or hyphen.* Additionally stopwords need to be eliminated.

1. Create an inverted index for the corpus. The index should contain only the document postings. What is the time taken to create this index?

2. Write a function to query against the index for a single token. On an average what is the query time for a single word query?

3. Extend the index created in the first question to contain the offset of the terms in the vocabulary in each document.

4. Use your new index to answer queries like *very simple*, *total failure* etc. How do you answer a phrase queries e.g. *interesting places to visit in London*

5. Extend your IR system to answer single wild card queries of the form $X*Y$ or $*X$ or $X*$.