Information Retrieval
Sessional - 6

September 15, 2023

0.1 Introduction

For these exercise we shall use the BBC Sports corpus. In Python we can open files for writing by using the following syntax:

```
fp = open(filename, "w")
fp.write(str)
fp.close()
```

0.2 Assignments

1. Try to create an inverted index for the corpus where the dictionary shall reside on the primary memory but the individual postings list shall be written out to the hard disk. More concretely the steps are as follows:
   (a) Parse and tokenize the corpus.
   (b) For each token:
      i. If the token exists open the corresponding file and append the docid.
      ii. If the token does not exist then open a new file and insert the doc id into the file. Maintain a link to the file in the dictionary which contain the token.
   (c) To answer a query read the postings for the corresponding term into the main memory and then call intersect.
2. Ingeneral how faster or slower is this approach compared to the *in primary approach* that we have followed so far.
3. Can we improve the efficiency of the method even further, if we maintain a single large postings file instead of opening multiple such files as discussed previously. Re implement the inverted index consisting of a single file now and answer queries using it.