

CMPE493: Information Retrieval Assignment 1

Elvan Karasu

1 Introduction

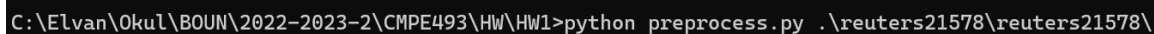
In this assignment, a simple search system for phrase and proximity queries is implemented using the Python Standard Library. [Reuters-21578](#) data set is used and it contains 21578 documents. [1] Raw data is preprocessed and inverted index is built. A query processor is implemented such that it finds the document IDs containing the queries using the inverted index.

2 Preprocessing

For normalization, punctuations and numbers are removed. The corpus contains the same amount of tokens (2567366) before and after case folding. 67363 unique tokens before case folding and 50651 unique tokens after case folding. The top 100 most common terms after case folding are: 'the, of, to, in, and, said, a, for, mln, it, dlrs, on, pct, is, that, its, from, by, will, vs, be, at, with, was, year, billion, he, us, has, as, an, would, cts, company, not, inc, net, which, bank, new, but, are, this, have, corp, were, last, market, had, stock, loss, or, shares, also, one, about, they, up, share, reuter, trade, been, two, shr, co, oil, may, sales, debt, more, first, banks, april, after, government, march, exchange, than, other, over, prices, group, dlr, profit, price, no, per, their, international, rate, foreign, ltd, interest, some, told, agreement, if, we, years, could'.

3 Inverted Index

A dictionary is built such that its keys are the unique tokens in the corpus. Value of each word is a list which has the first element as the document frequency (number of documents containing that term). The second element of the list is another dictionary which has document IDs as keys. The values are lists such that the first element of the list is term frequency (number of occurrences of the word in the document). The second element is the postings list for the occurrences of the word in that document. The screenshots for the running of indexing and query modules are given in Figure 1-3 below.



```
C:\Elvan\Okul\BOUN\2022-2023-2\CMPE493\HW\HW1>python preprocess.py .\reuters21578\reuters21578\
```

Figure 1: Command for running the indexing module

```
C:\Elvan\Okul\BOUN\2022-2023-2\CMPE493\HW\HW1>python query.py term_dict.pkl "old crop cocoa"
1
```

Figure 2: Command for running a phrase query

```
C:\Elvan\Okul\BOUN\2022-2023-2\CMPE493\HW\HW1>python query.py term_dict.pkl old 1 cocoa
1 19570
```

Figure 3: Command for running a proximity query

References

- [1] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>