CmpE 493

Information Retrieval

Assignment 1 Report

Atakan Arıkan – 2011400243

1. Preprocessing the data

`readStopwords(){`

Reads the data written in the "stopwords.txt" file, and puts them in `HashSet<String> stopWords`.

`}`

`readArticles(){`

Starts reading the articles from reuters21578/reut2-000.sgm line by line. If the line contains "<REUTERS>" tag, starts putting the following lines into the variable "article" until it sees "</REUTERS>" tag in a line. Then it extracts the article to be indexed from the article variable. There are 3 cases:

```
if (article.contains("</BODY>")) { // article has a body tag, normal case
} else if (article.contains("</TITLE>")) { // article has a title tag, but not a body tag
} else { // no body neither title tag
}
```

Then, extracts the article to be indexed, tokenizes it, gets rid of the stopwords and finally, stems each of the tokens inside the article in this line:

String[] tokens = stemmed(deleteStopWords(tokenize(tobeIndexed))).split(" ");

Then, updates or creates an entry in *HashMap<String, String> invertedIndex* and updates *HashMap<String, Integer> wordFrequency*

}

`tokenize(String line){`

It's a simple tokenizer, splits the given line by whitespace. Then for each token, if the token is not convertible to the double, iterates through each character and gets rid of the nonword characters. For instance, U.S.A. becomes USA.

(a) How many tokens does the corpus contain before stopword removal and stemming?

> 3233635

(b) How many tokens does the corpus contain after stopword removal and stemming?

> 2599137

(c) How many terms (unique tokens) are there before stopword removal, stemming, and case-folding?

> 134824

(d) How many terms (unique tokens) are there after stopword removal, stemming, and casefolding?

> 67343

(e) List the top 20 most frequent terms before stopword removal, stemming, and casefolding?

> [&#3;, 19780] [of, 15542] [the, 15097] [Reuter, 14361] [and, 14237] [to, 14164] [said, 13564] [in, 12888] [The, 11048] [for, 10917] [it, 9382] [mln, 8342] [on, 8248] [said., 8033] [its, 7917] [is, 7841] [from, 7605] [by, 7477] [with, 7230] [at, 7201]

(f) List the top 20 most frequent terms after stopword removal, stemming, and case-folding?

> [reuter, 19058] [said, 15465] [to, 14879] [on, 10055] [dlr, 8882] [mln, 8609] [from, 7798] [by, 7692] [at, 7552] [year, 6596] [pct, 6269] [compani, 6181] [ha, 6163] [that, 6092] [inc, 6014] [corp, 5388] [not, 4758] [new, 4558] [would, 4235] [share, 4127]

2. Data Structures

The DS used for the inverted index is in the following format. It's a HashSet, for its convenient lookup time.

<String, String>

<word1> --> <docid1>, <docid2>, <docid3>, <docid4>...

<word2> --> <docid1>, <docid2>, <docid3>, <docid4>...

...