Cyprus International University

CMPE501, ISYE501, CMPE611, MISY501

Advanced Programming Languages

Project: Document Retrieval System

Weight of the project: %30

Due Date: 22 June 2020

The aim of this project is to implement a simple text retrieval system. You should implement a simple UI that gets a search keyword from the user and shows the matching documents with title, date and document body (content). Please follow the steps below to complete the task.

Part A. (75pts) Implement a simple text retrieval system

1) Download the text archive (corpus) from the following URL:
   http://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/
   Extract the content of reuters21578.tar.gz
   Consider only the files with .sgm extension that are text files. Each .sgm file contains many documents.
2) Implement an application to read `.sgm` files from `reut2-000.sgm` to `reut2-020.sgm` and to parse the documents. You should implement a `Document` class in Java that will contain DATE, TOPICS, TITLES and BODY of the document. Your parser will read the `.sgm` file and split the `.sgm` file into Document class.
3) Store your Document in a Vector or ArrayList (Collecttion)
4) Implement an UI to get a search Keyword and search the keyword from the Collection.
5) Show the documents in a `TextArea` that contains the given keyword. Your results should be formatted to show title, date and body (content) of each matching document.

Part B. (25pts) Extend your search engine.

1. Update your document retrieval system to make the search in parallel with threads. You should have 4 threads to search the documents.
2. Implement a stemmer to find more meaningful documents with text search. You can use `PorterStemmer` algorithm to retrieve documents that contains *meet*, even if you search for *meeting* or *meets*. You can use any existing library from the internet like:
   URL https://tartarus.org/martin/PorterStemmer/java.txt

Part C. BONUS: (Additional %20)

You can use sequential search that checks all documents one by one for this project. However, if you use any Tree structure or sorted inverse document structure to organize the documents and make faster

search this will be considered as bonus grades. Please specify this in your document if you have implemented this part.

Deliverables:

1) Only moodle upload is accepted until the due date. Sending email or weetransfer etc is not accepted.
2) All deliverables must be submitted as a single zip file. You should name your zip file with your student number and name.
3) Your source code: (Your source code copied from your project folder) All `.java` files and any other resources you have used must be included.
4) Documentation: A pdf file that explains your project, solution approach and contains screenshots etc. Your documentation should have a cover page with your name, student number and department.
5) Any similar work among students and codes downloaded from the internet will not be considered and will be not be marked.

Example

1) Consider the file: `reut2-000.sgm`
2) Your parser should retrieve the following text as the second Document

```
<REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5545" NEWID="2">
<DATE>26-FEB-1987 15:02:20.00</DATE>
<TOPICS></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F Y
&#22;&#22;&#1;f0708&#31;reute
d f BC-STANDARD-OIL-&lt;SRD>-TO   02-26 0082</UNKNOWN>
<TEXT>&#2;
<TITLE>STANDARD OIL &lt;SRD> TO FORM FINANCIAL UNIT</TITLE>
<DATELINE>    CLEVELAND, Feb 26 - </DATELINE><BODY>Standard Oil
Co and BP North America
Inc said they plan to form a venture to manage the money market
borrowing and investment activities of both companies.
    BP North America is a subsidiary of British Petroleum Co
```

```
Plc &lt;BP>, which also owns a 55 pct interest in Standard Oil.
    The venture will be called BP/Standard Financial Trading
and will be operated by Standard Oil under the oversight of a
joint management committee.

 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

3) Your Document Class should store this data in a as in the following:
   Date: 26-FEB-1987 15:02:20.00
   Title: STANDARD OIL <SRD> TO FORM FINANCIAL UNIT
   Body: Standard Oil Co and BP North America Inc said they plan to form a venture to manage the money market borrowing and investment activities of both companies. BP North America is a subsidiary of British Petroleum Co Plc <BP>, which also owns a 55 pct interest in Standard Oil.
   The venture will be called BP/Standard Financial Trading and will be operated by Standard Oil under the oversight of a joint management committee.