Ramy El Khayat
Rabih Talaba
Instructor: Cherie Ding
CPS 842
September 28, 2016

Assignment 1 Report

The purpose of this assignment is to perform search on a collection of documents and obtain information each document such as the ID number , title , position of word and a brief summary of the document containing the term searched.

The search is conducted by the use of a dictionary and a postings list after parsing the collection into documents and then parsing the document to generate terms and their frequencies.

Java was used for the implementation.

Files breakdown

Document: A document object which retains info about the documents attributes (ID, date, title, abstract,etc)

DocumentParser: The parser has an instance of a document that is used select a document to be parsed. The parsing process is as follows a method called findTerms returns a set of all terms in the document in this case the terms we care about are in the abstract and title together they form a full document also a Stemmer is available to stem each word if the user wants to use stemming. The docParser also calculates frequencies for every term and the position of every word in the doc.

FileParser: this object parses the collection in “cacm.all” and divides the collection into document objects by using a scanner and identifying the flags in the collection for IDs , titles and abstracts we were able to parse the collection.

DocumentFrequency: a wrapper class to update frequencies

Stemmer: implements porter’s stemming algorithm by eliminating certain suffixes from words to obtain the stem of each word (stemming is optional at the beginning of the program)

Posting: a posting contains about the term frequency , the ID of document where it occurs and a set of its positions within that document.

Invert: creates the dictionary and postings list files

Test: uses the files created by Invert to locate terms in the collection.

Data Structures

The dictionary and postings list are made using hash maps the dictionary data structure was simple enough to just have the term and its document frequency

```
private Set<String> dictionary;
private Set<Integer> positions;
private Map<Integer, Document> documents;
private Map<String, DocumentFrequency> documentFrequencies;
```

The document frequencies is the dictionary as it counts the frequency of every word using the Document frequency class.

```
private static HashMap<String,ArrayList<Posting>> postingsList= new  HashMap<String,ArrayList<Posting>>();
```

Positions is the set of positions for every term (calculate by the docParser's method parsePositions).

Postings List is a hash map of String (key) and the value is an array of Posting objects containing the information needed to retrieve terms from the document collection

How to run the program

1. Place all files into one folder
2. in terminal navigate to the folder and compile the following command: `java *.java`
3. run the file Test using the command `java Test`
4. Wait for the program to create the postings then it will prompt you to enter a search term