COMP4321 Lab 3 - Stemming and Stopword Removal

1 Introduction

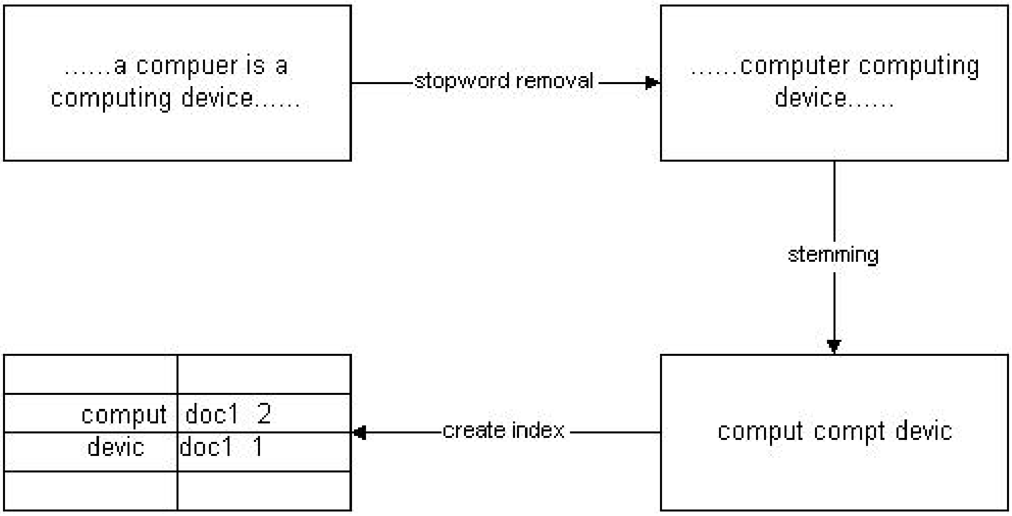
A search engine usually provides a friendly user interface that you can input the keywords you want to search for. Based on your input query, the search engine can retrieve all the documents that are related to the query according to its own retrieval algorithm and return a result list to you. A concrete example of a search engine is the one you used in our library.

Basically nearly all the search engines can accept keywords search; others even permit users to input questions or support languages other than English. The objective of a search engine is to provide the most convenient and effective search mechanism to its users. Its result lists should contain documents which really conform to the user's original interest.

2 Working Mechanism of a Simple Search Engine

Now we will introduce a mechanism of a simple search engine which is based on the vector space model.

- Stopword removal: Words which are very frequent and do not carry meaning (such as "a", "the") are called stopwords. These words are assumed not to carry any important information and so are usually ignored in order to save storage space of the inverted file. First, you should define the list of stopwords. Then, when you read in a new document you should remove all the stopwords before proceeding to the next stage.
- Stemming: In English, words having the same stem are usually assumed to have similar meaning, such as "computing", "compute", "computed" and "computation". In order to improve the recall of the search (i.e., to get relevant documents which don't contain the exact words as specified in the query), stemming is performed to remove the affix and successor. Porter's algorithm is a well-known stemming algorithm.
- Creating the index: You should maintain the inverted index that contains the word/posting pairs. If a word has already existed in the index file, you can store the record containing the id of this document and the term frequency to its posting. Otherwise you should create a new key and store the corresponding postings in the posting list.


......a compuer is a computing device......
stopword removal
......computer computing device......
stemming
comput compt devic
create index
comput doc1 2
devic doc1 1

The above steps are used to create the index file. After we have such kind of index, we can provide the search service. Actually there are many different search algorithms, each with its own advantage sand disadvantages. We may adopt the simplest $tf * idf/max_t f$ weighting scheme and assume that all of the keywords that users input have the same weight.

- Keywords preprocessing: When a user inputs the keywords, we should preprocess them. First removing all the stopwords and then doing the stemming.
- Searching: Now we can do the search according to the keywords. Basad on the ranking algorithm, we compute the score of all the relevant documents and rank them.
- Returning the result: According to the ranking, we can return the result documents list to the user.

3 Stemmer and Stopword Remover

In the project, you are recommended to use the "Porter stemming algorithm". It is fast and has a performance comparable to other complex algorithms. Porter.java is a Java implementation of the Porter stemming Algorithm, which was implemented by Fotis Lazarinis.

For stopword removal, there are many stopword removers ready for use. However, most of them are C programs, which are unnecessarily complicated. You are recommended to use a simple data structure in java to help you perform this task efficiently.

StopStem.java is an example for stopping and stemming in Java. To compile it, you need to firstly place the Porter.java into a directory named "IRUtilities" and compile the Porter.java into Porter.class.

4 Exercise

You may notice that in the StopStem.java, the stopword list is hard-coded by inserting each word line by line. You are required to modify the example program to read stopwords from a file, stopwords.txt.

You only need to modify the code in "StopStem" function, which is used to read the stopwords from txt file into memory. You are not required to touch other code in sample program. Thus, a package named as "IRUtilities" should be created under project and Porter.java should be put under the package.

Learn by yourself about how to read external txt file by "BufferedReader".

Submission
no need to submit