**INFORMATION RETRIEVAL**

**Submitted By:**

Snehith Desu (U00768054)

**Objective:**

The aim of the project is to design and implement Search Engine. In Phase I, we have built the indexing component, which will take cran.all text file and parse it into 1400 text files to produce a searchable, persistent data structure. In Phase II, a query is entered to display all the text files related to it.

**Design Architecture:**

**Parsing:**

In Phase-1, a large data set cran.all is parsed into 1400 text files. Using .I as the delimiter, the corpus file is parsed into 1400 files.

* The data between two consecutive .I is considered as a single file.

**Tokenization:**

Each individual document is converted into a set of keywords or tokens.

* In this step, all the special characters, numbers are removed, trimmed and then converted to lower case.
* We use String Builder, String Tokenizer classes to tokenize all the documents into tokens

**Stemming:**

All the inflected words are reduced to their root words to enhance the efficiency of the retrieval system. In this step we use Porter’s Stemming Algorithm. (http://ir.dcs.gla.ac.uk/resources/linguistic\_utils/porter.java)

**Stop Word Elimination:**

These tokens are compared with the list of stop words. If the tokens are present in the list of stop words, those words are eliminated. (Stop words from the link <http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop>)

**Indexing:**

The resultant tokens (after Stemming) are stored in a Hash Map with Token, Document Id Term Frequency (Count).

**Data Structures Used:**

Used Array List, Hash Map, String Builder.

* Array List is used for storing the Stop Words.
* Hash Map is used for storing Tokens.
* String Builder for storing file temporarily (for tokenizing the file).

**Benchmarks for Phase-I:**

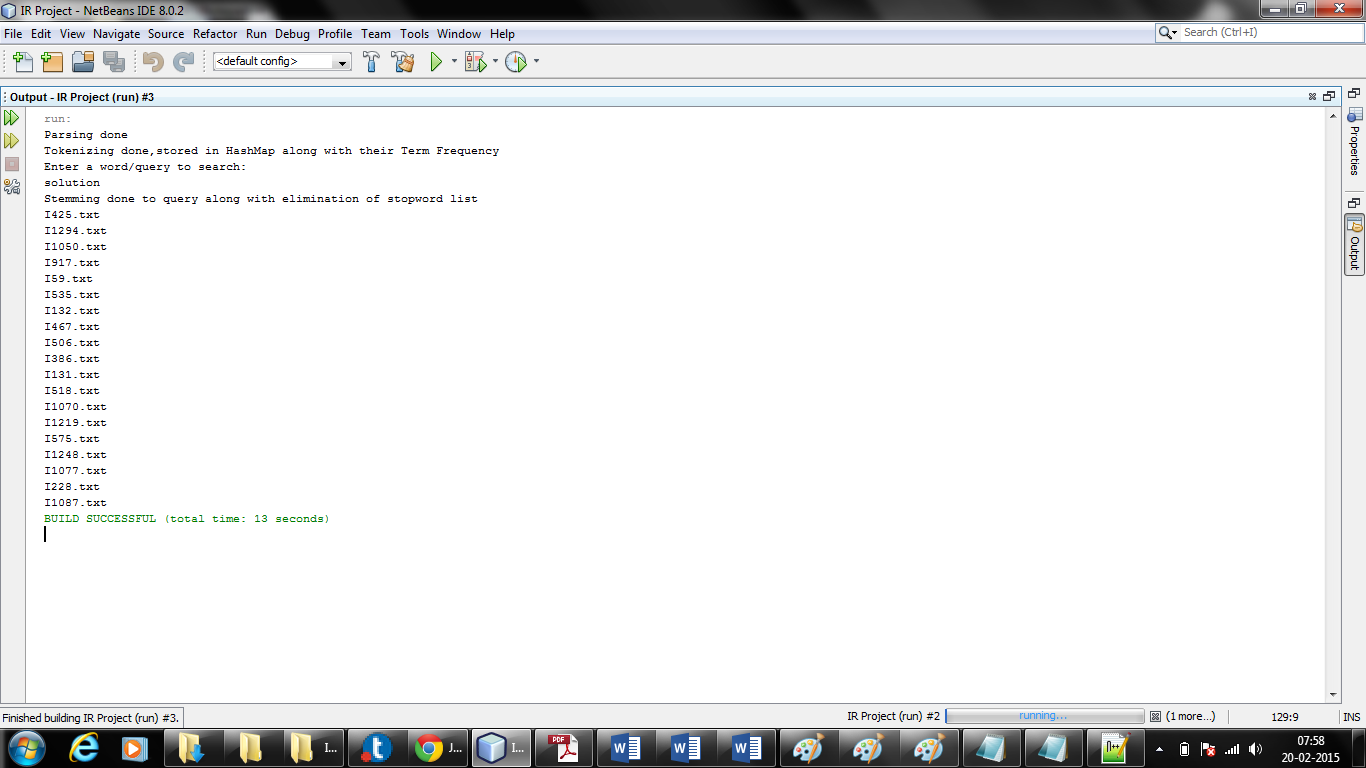
1. **Corpus Statistics:**
   1. There are 1400 documents in the collection
   2. There are 262,912 total number of words and 7242 unique words in the collection.
   3. There are 6 inverted file entries. Longest posting list length is 830 documents per term and shortest posting list length is 1 document per term. Average posting list length is 151.488 documents per term.
2. **Inverted Collection:**
   1. It took 5 seconds to index the collection.
   2. 2.71MB of the total disk space is required for index and data structures.
   3. Size of index relative to size of corpus dataset is 1.70MB

**Phase-II:**

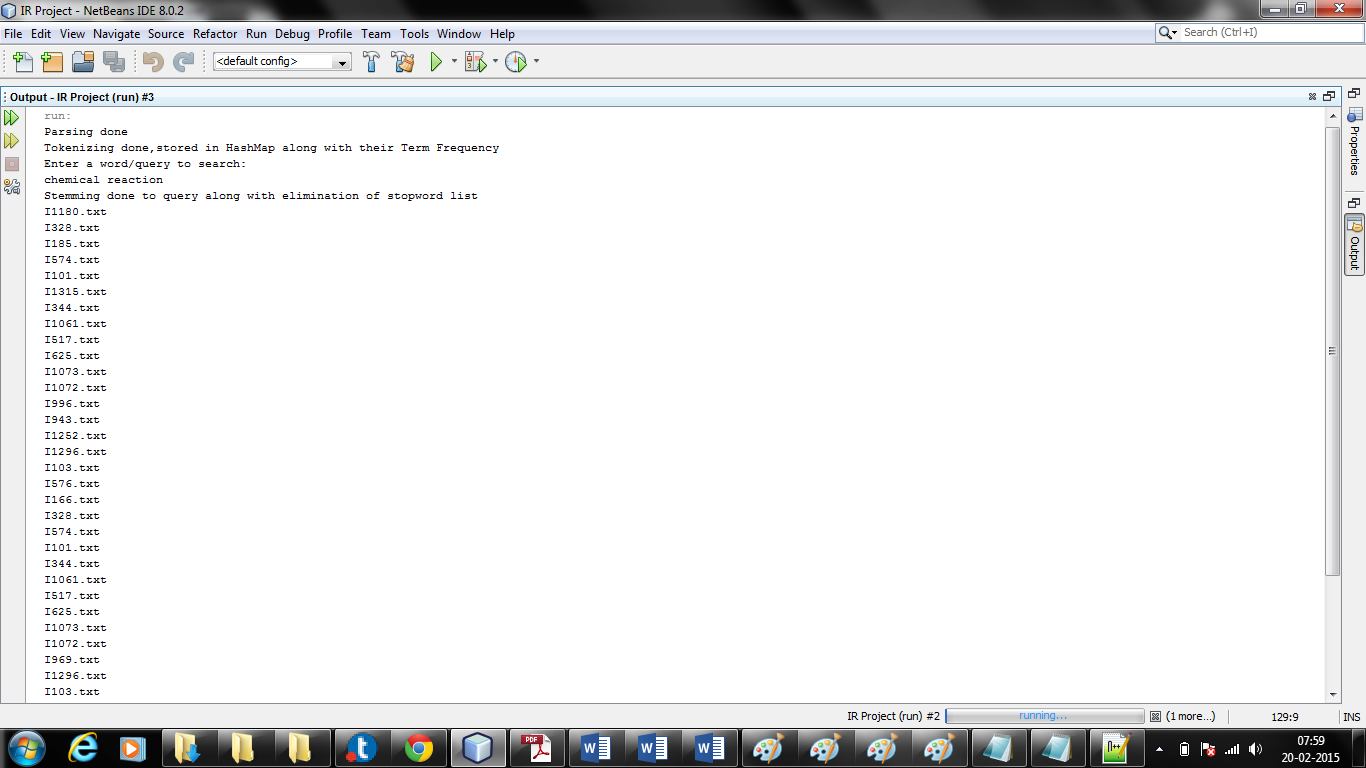
In Phase-II user enters a Token/Query in the Console Window and retrieves the relevant records. Top 20 documents for every Token are displayed to the user. If there exists a stop word in the query then no document will be retrieved for the corresponding stop word.

**Benchmark for Phase-II:**

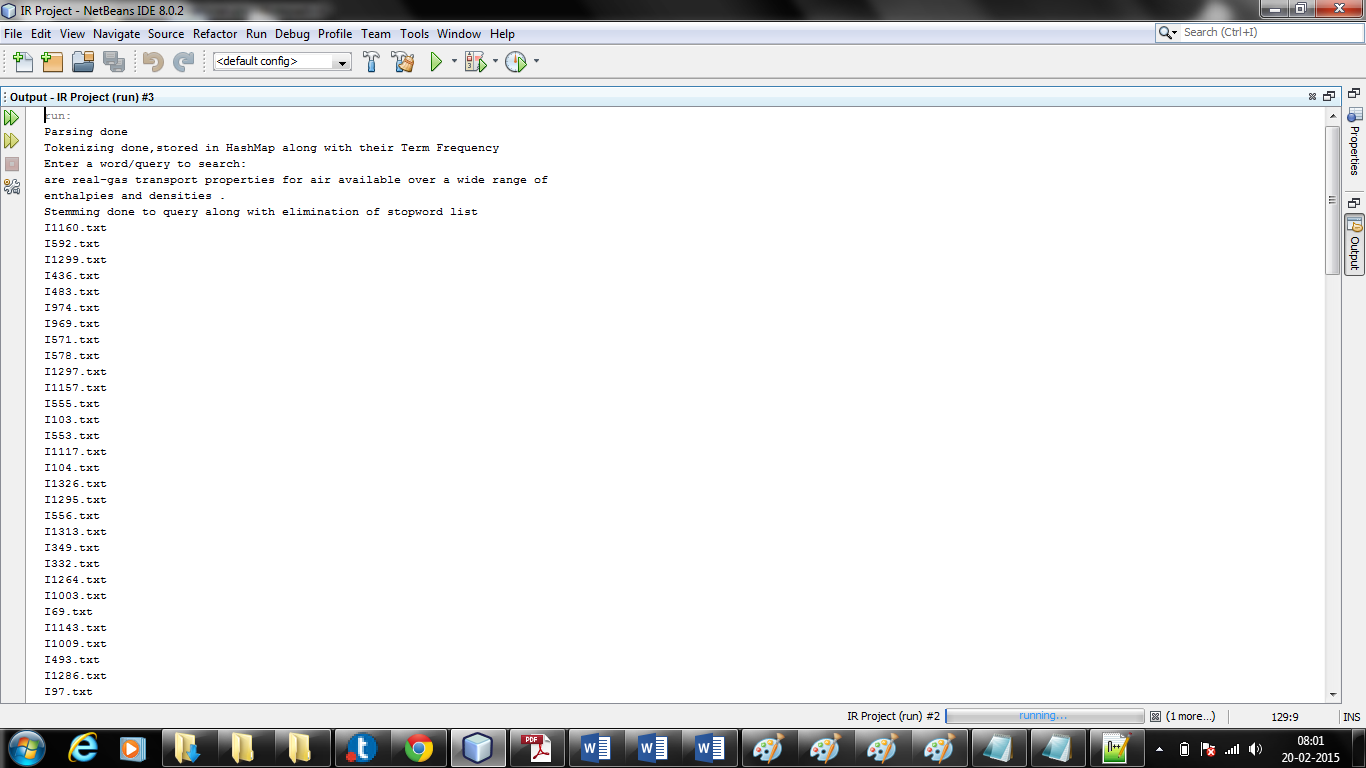
* Single word as query, the output observed is shown in the screenshot below.



* With two words as query, the output observed is shown in the screenshot below.



* Giving a Phrase as query, the output observed is shown in the screenshot below.



**Precision and Recall:**

Precision: [{Relevant Documents} ∩ {Retrieved Documents}] / {Retrieved Documents}

Recall: [{Relevant Documents} ∩ {Retrieved Documents}] / {Relevant Documents}

|  |  |  |
| --- | --- | --- |
| Query | Precision | Recall |
| Are real gas transport properties for air available over a wide range of enthalpies and densities | 17/20 = 0.85 | 17/150 = 0.113 |
| What is the basic mechanism of the transonic aileron buzz | 15/20=0.75 | 15/200=0.075 |