Hyun Han / 160152230

**Code description:**

The information of documents is fixed. Therefore the functions related to documents have been written in the constructor to minimize the running time. For each term weighting scheme, different functions about a given query and cosine similarity with documents has been written.

**The extent of implementation achieved (all succeeded) :**

Evaluation of the performance of the system over the CACM test collection under alternative configurations with different term weighting scheme used during retrieval has been undertaken. The result is below

**Performance:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Binary** | None(1.47s) | Stoplist(1.04s) | Stemming(1.47s) | Both(1.03s) |
| Rel\_Retr | 44 | 84 | 59 | 104 |
| Precision | 0.07 | 0.13 | 0.09 | 0.16 |
| Recall | 0.06 | 0.11 | 0.07 | 0.13 |
| F-measure | 0.06 | 0.12 | 0.08 | 0.14 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TF** | None(1.47s) | Stoplist(1.03s) | Stemming(1.54s) | Both(1.07s) |
| Rel\_Retr | 50 | 105 | 72 | 123 |
| Precision | 0.08 | 0.16 | 0.11 | 0.19 |
| Recall | 0.06 | 0.13 | 0.09 | 0.15 |
| F-measure | 0.07 | 0.15 | 0.10 | 0.17 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TFIDF** | None(4.82s) | Stoplist(2.90s) | Stemming(4.51s) | Both(3.05s) |
| Rel\_Retr | 132 | 140 | 166 | 172 |
| Precision | 0.21 | 0.22 | 0.26 | 0.27 |
| Recall | 0.17 | 0.18 | 0.21 | 0.22 |
| F-measure | 0.18 | 0.19 | 0.23 | 0.24 |

**Analysis:**

In each of the methods, using both stoplist and stemming obviously return the best result, and stoplist, stemming and none in order. Removing non-content words are turned out to be effective. Considering F-measure stoplist is more effective than stemming. But in TFIDF stemming F-measure is higher in stemming tan stoplist.

Comparing across the different term weighting method, there are no significant gap between binary and TF. However the all of the figures dramatically increase in TFIDF compare to the other two.