CS657: Information Retrieval

Ankush Singh 150107

06-02-2018

Assignment 1

In the assignment, I have used Terrier as my tool. In Terrier, the stemmers I have used are : PorterStemmer, WeakPorter and English Snowball.

AND queries and OR queries have been separately dealt -each indexed first without stopword removal and then with stopword removal.
The following data has been recorded in tables for each of the 12 cases.
1. Time-taken to generate index
2. Size of the vocabulary.
3. Size of retrieved set of documents.
4. Precision, Recall and F-score of all query groups.

Size of corpus = 125586. Maximum Retrieval Limit=50000

Analysis/Comments

- Porter Stemmer is the fastest stemmer(out of the three) followed by Weak-Porter Stemmer and then EnglishSnowball
- WeakPorter Stemmer makes the largest vocabulary during indexing though the sizes are almost comparable for all stemmers.
- AND queries were more difficult and had rarer results than OR queries
- OR queries suffered in Precision and F-score because of larger retrieval bu t had an astounding recall.
- Weak Porter performed best in precision as well as F-score followed by English Snowball which was best in Recall while Porter Stemmer always struggled.
- Since F-score is a more accurate measure than precision or recall, Weak Porter Stemmer claims the best stemmer in evaluation though it had a large vocabulary and a slow pace.
- Stopword removal had very negligible effect on evaluation measures. Also, it made indexing slow.

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	110.306	113.802	162.685
AND	Removed	112.859	115.889	137.858
OR	Not Removed	69.44	68.05	90.385
OR	Removed	68.252	95.513	118.844

Table 1: Retrieval Statistics : Time-taken to index (in seconds)

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	191156	207079	191930
AND	Removed	191156	207079	191930
OR	Not Removed	191156	207079	191930
OR	Removed	191156	207079	191930

Table 2: Retrieval Statistics : Size of Vocabulary

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	802	579	800
AND	Removed	802	579	800
OR	Not Removed	50000	50000	50000
OR	Removed	50000	50000	50000

Table 3: Retrieval Statistics :Size of retrieved set

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	0.197	0.246	0.198
AND	Removed	0.197	0.246	0.198
OR	Not Removed	0.0303	0.029	0.0304
OR	Removed	0.0303	0.029	0.0304

Table 4: Evaluation Measures : Precision

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	0.242	0.219	0.244
AND	Removed	0.242	0.219	0.244
OR	Not Removed	0.9954	0.9953	0.9947
OR	Removed	0.9954	0.9953	0.9947

Table 5: Evaluation Measures : Recall

		Stemmer Type		
Query Type	**Stopword Status**	**Porter Stemmer**	**WeakPorter Stemmer**	**EnglishSnowball Stemmer**
AND	Not Removed	0.217	0.232	0.219
AND	Removed	0.217	0.232	0.219
OR	Not Removed	0.05877	0.057	0.05899
OR	Removed	0.05877	0.057	0.005899

Table 6: Evaluation Measures : F-score