Data Mining Technology for Business and City

Siddhant Tandon 1771650

Syed Hassan Abbas 1760140

1.A simple statistic on the used dataset: #documents and #queries.

Collection name: cran

#Documents: 1400

#Queries: 222

```
siddhant@siddhant-P6661-MD60196:~/DMT/Homework_1_dmt$ more cran-{text,title}.properties
::::::::::::::
cran-text.properties
::::::::::::::
documents=1400
terms=7471
postings=122934
maxcount=100
indexclass=it.unimi.di.big.mg4j.index.QuasiSuccinctIndex
skipquantum=256
byteorder=LITTLE_ENDIAN
termprocessor=it.unimi.di.big.mg4j.index.DowncaseTermProcessor
batches=1
field=text
size=2790772
maxdocsize=662
occurrences=226675
::::::::::::::
cran-title.properties
::::::::::::::
documents=1400
terms=1800
postings=15742
maxcount=6
indexclass=it.unimi.di.big.mg4j.index.QuasiSuccinctIndex
skipquantum=256
byteorder=LITTLE_ENDIAN
termprocessor=it.unimi.di.big.mg4j.index.DowncaseTermProcessor
batches=1
field=title
size=228551
maxdocsize=43
occurrences=16619
```

Fig 1. Screenshot: inverted index using default stemmer

2. A list of used stemmers.

- Default Stemmer
- English Stemmer
- English Stemmer able to filter stop words

3. A list of used scorer functions.

- Count Scorer
- BM25
- TF-IDF

4. The script to create the collection.

find Cranfield_DATASET/cran -iname *.html | java it.unimi.di.big.mg4j.document.FileSetDocumentCollection -f HtmlDocumentFactory -p encoding=UTF-8 cran.collection

*all the html files were in the Cranfield_DATASET/cran folder

5. The scripts to create inverted indexes (all stemmers)

Collection using default stemmer:
java it.unimi.di.big.mg4j.tool.IndexBuilder --downcase -S cran.collection cran

Collection using English Stemmer:
java it.unimi.di.big.mg4j.tool.IndexBuilder -t it.unimi.di.big.mg4j.index.snowball.EnglishStemmer -S cran.collection cran

Collection using English Stemmer able to filter stopwords:
java it.unimi.di.big.mg4j.tool.IndexBuilder -t homework.EnglishStemmerStopwords -S cran.collection cran

6. The scripts to obtain the results from the search engine (for all scorer function).

Default Stemmer:

Scorer: TF-IDF
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "TfIdfScorer" "text_and_title" output_cran__defaultStemmer__TfIdfscorer.tsv

Scorer:BM25
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "BM25Scorer" "text_and_title" output_cran__defaultStemmer__BM25Scorer.tsv

Scorer: Count Scorer
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "CountScorer" "text_and_title" output_cran__defaultStemmer__Count_scorer.tsv

English Stemmer:

Scorer: TF-IDF
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "TfIdfScorer" "text_and_title" output_cran__EnglishStemmer__TfIdfScorer.tsv

Scorer: BM25Scorer
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "BM25Scorer" "text_and_title" output_cran__EnglishStemmer__BM25Scorer.tsv

Scorer: Count Scorer
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "CountScorer" "text_and_title" output_cran__EnglishStemmer__CountScorer.tsv

English stemmer able to filter stop words:

Scorer:TF-IDF
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "TfIdfScorer" "text_and_title" output_cran__EnglishStemmer_stopwords__TfIdfScorer.tsv

Scorer:BM25
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "BM25Scorer" "text_and_title" output_cran__EnglishStemmer_stopwords__BM25Scorer.tsv

Scorer:CountScorer
java homework.RunAllQueries_HW "cran" ./cran_all_queries.tsv "CountScorer" "text_and_title" output_cran__EnglishStemmer_stopwords__CountScorer.tsv

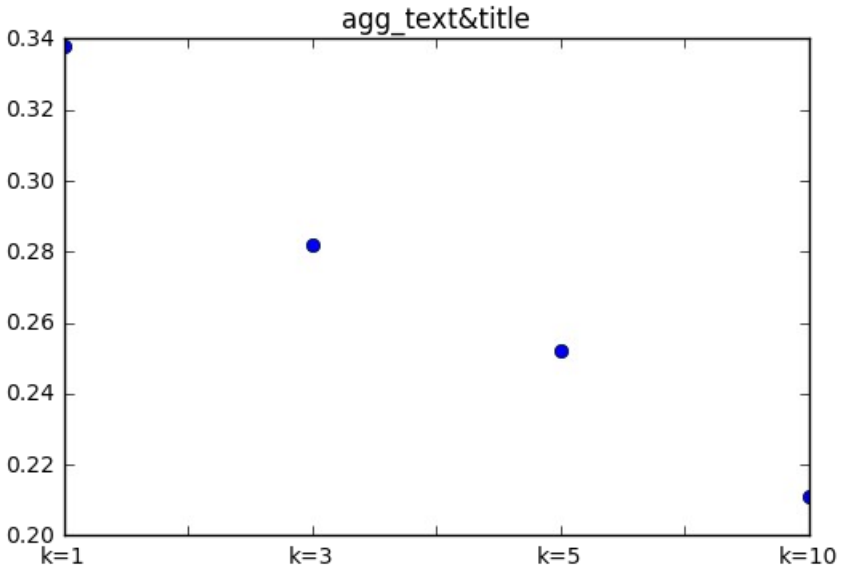
7. The Average R-Precision for each stemmer-scorer_function configuration end for the aggregation algorithm: (3X3 + 1)= 10 Average R-Precision values.

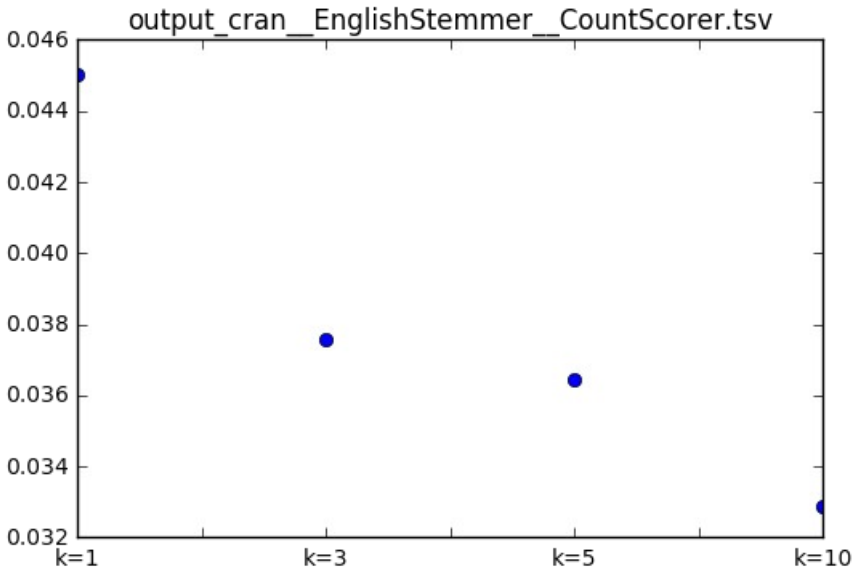
File name	Avg_R-Precision
output_cran__EnglishStemmer__CountScorer.tsv	0.0277202579
output_cran__defaultStemmer__Count_scorer.tsv	0.0244028171
output_cran__defaultStemmer__BM25Scorer.tsv	0.2549430381
output_cran__EnglishStemmer_stopwords__TfIdfScorer.tsv	0.1908839392
output_cran__defaultStemmer__TfIdfscorer.tsv	0.179327875
output_cran__EnglishStemmer_stopwords__CountScorer.tsv	0.1579409437
output_cran__EnglishStemmer_stopwords__BM25Scorer.tsv	0.2664739777
output_cran__EnglishStemmer__BM25Scorer.tsv	0.2623916658
output_cran__EnglishStemmer__TfIdfScorer.tsv	0.1893509075
agg_file_BM25_EngStemmerstopwords_text&title_k=5	0.2515342899

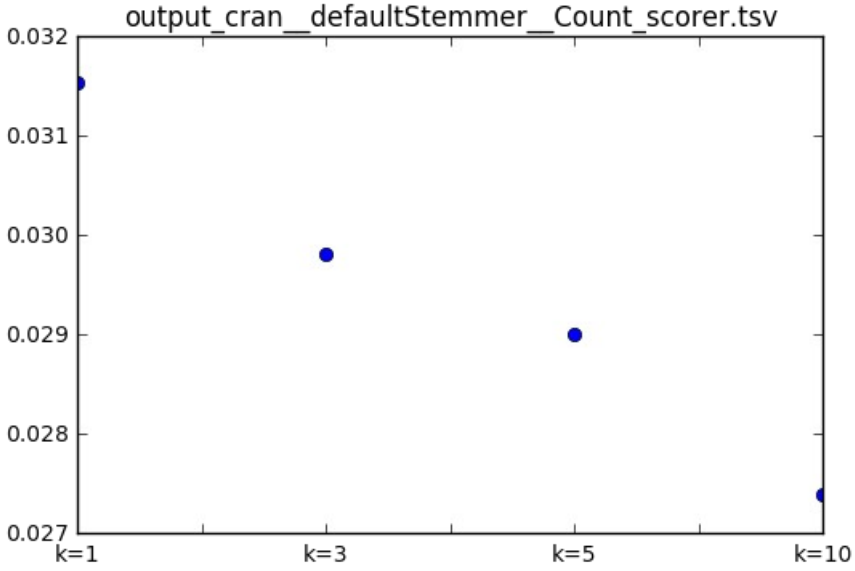
*the last score is for a default value of k=5. This file can be generated by running fagin.py and with a user input of k=5

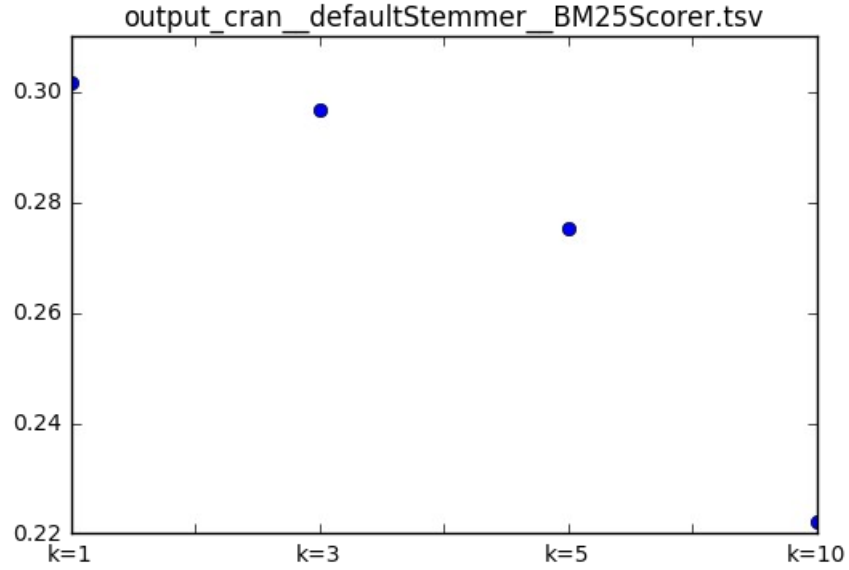
8. The plot of the average nMDCG:

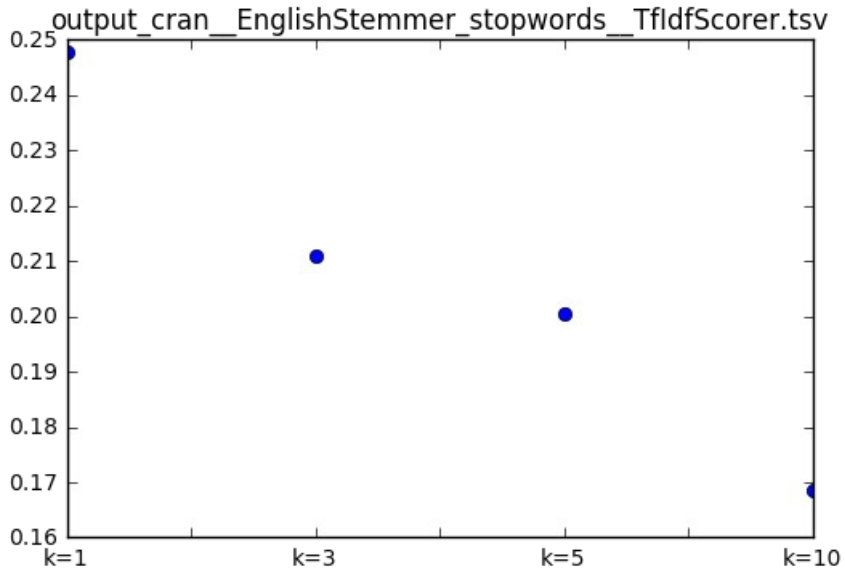
File name	k=1	k=3	k=5	k=10
agg_text&title	0.3378378378	0.2818279889	0.2522800969	0.2108548925
output_cran__EnglishStemmer__CountScorer.tsv	0.045045045	0.037564611	0.0364451361	0.0328989832
output_cran__defaultStemmer__Count_scorer.tsv	0.0315315315	0.0298193974	0.0290020524	0.0273912166
output_cran__defaultStemmer__BM25Scorer.tsv	0.3018018018	0.2968489587	0.2754096289	0.2222101296
output_cran__EnglishStemmer_stopwords__TfIdfScorer.tsv	0.2477477477	0.2108962548	0.2004971318	0.1685611579
output_cran__defaultStemmer__TfIdfscorer.tsv	0.2387387387	0.2086545619	0.1957670385	0.1655774937
output_cran__EnglishStemmer_stopwords__CountScorer.tsv	0.2432432432	0.2043336165	0.1832649964	0.1526850553
output_cran__EnglishStemmer_stopwords__BM25Scorer.tsv	0.3018018018	0.3173133475	0.2903511551	0.2434295144
output_cran__EnglishStemmer__BM25Scorer.tsv	0.3018018018	0.3379401767	0.2920079587	0.2379225695
output_cran__EnglishStemmer__TfIdfScorer.tsv	0.2567567568	0.2241449889	0.2069466633	0.1754884022

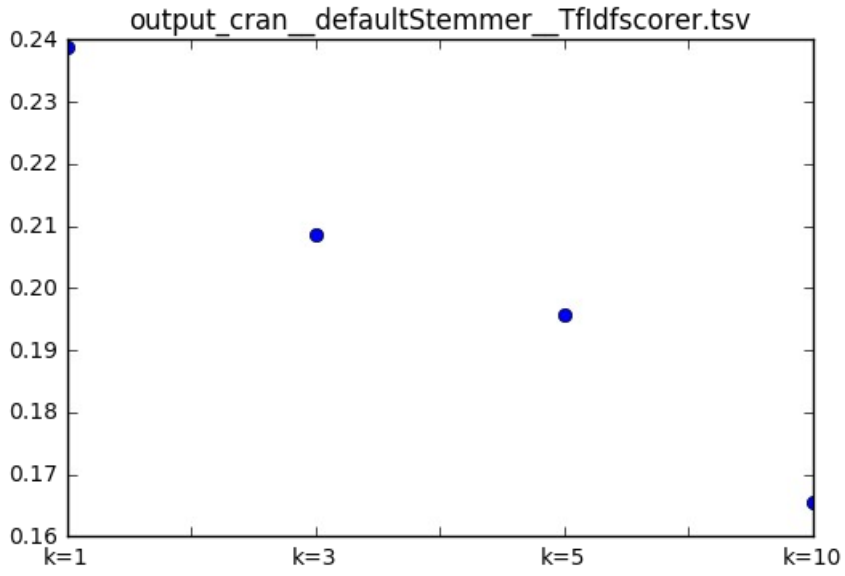


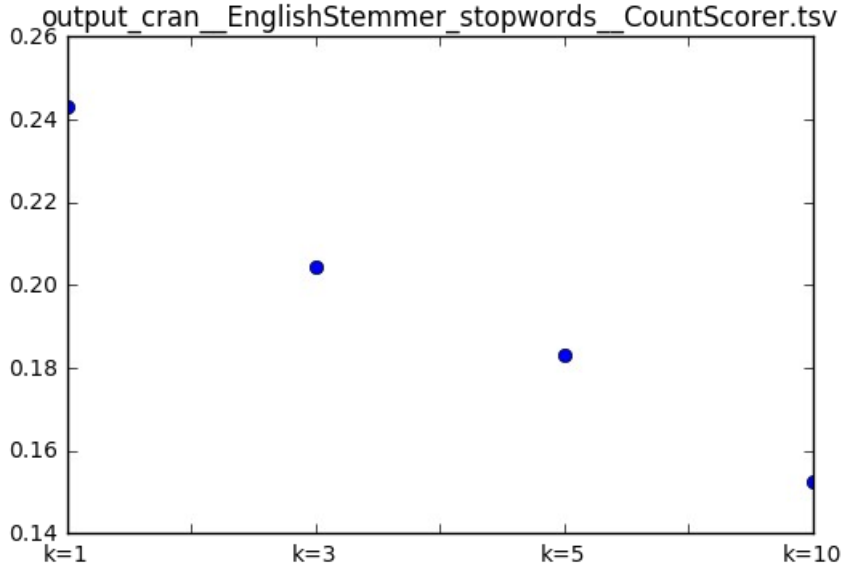

output_cran__EnglishStemmer__CountScorer.tsv
0.046
0.044
0.042
0.040
0.038
0.036
0.034
0.032
k=1
k=3
k=5
k=10

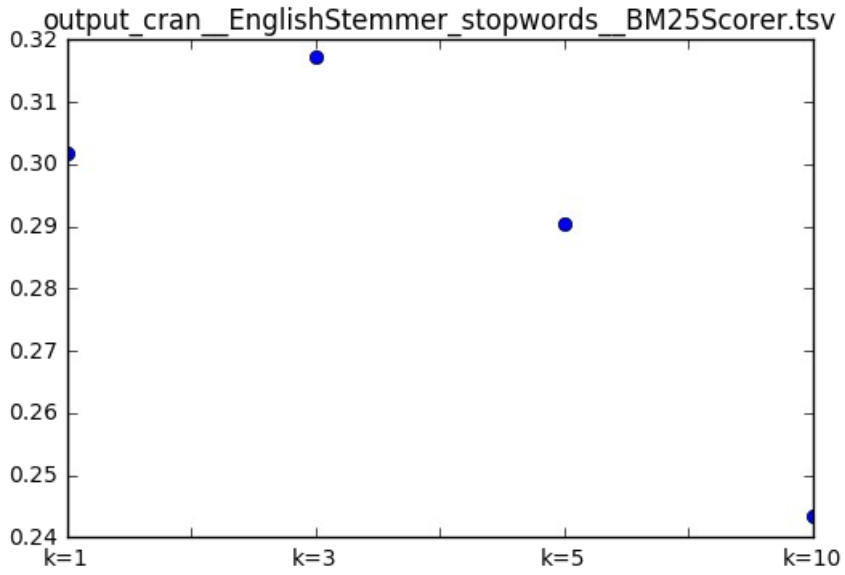

output_cran__defaultStemmer__Count_scorer.tsv
0.032
0.031
0.030
0.029
0.028
0.027
k=1
k=3
k=5
k=10

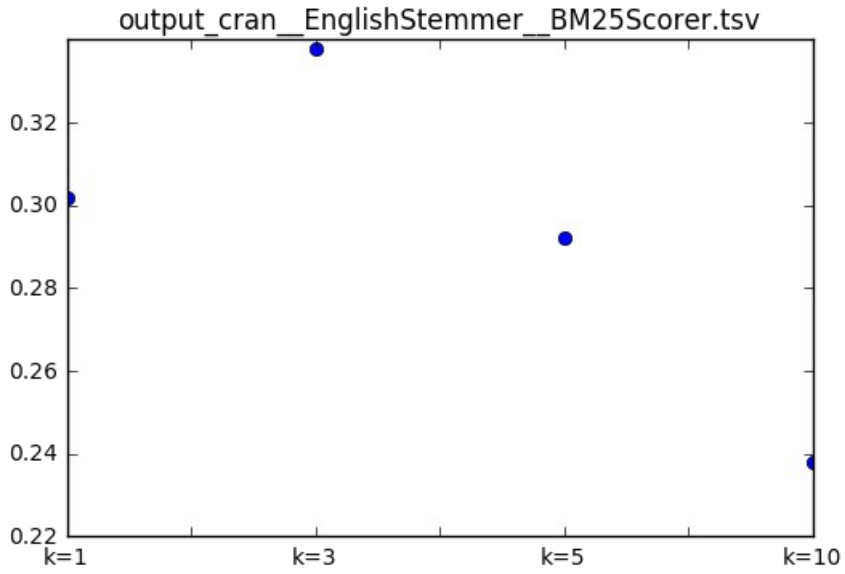

output_cran__defaultStemmer__BM25Scorer.tsv
0.30
0.28
0.26
0.24
0.22
k=1
k=3
k=5
k=10

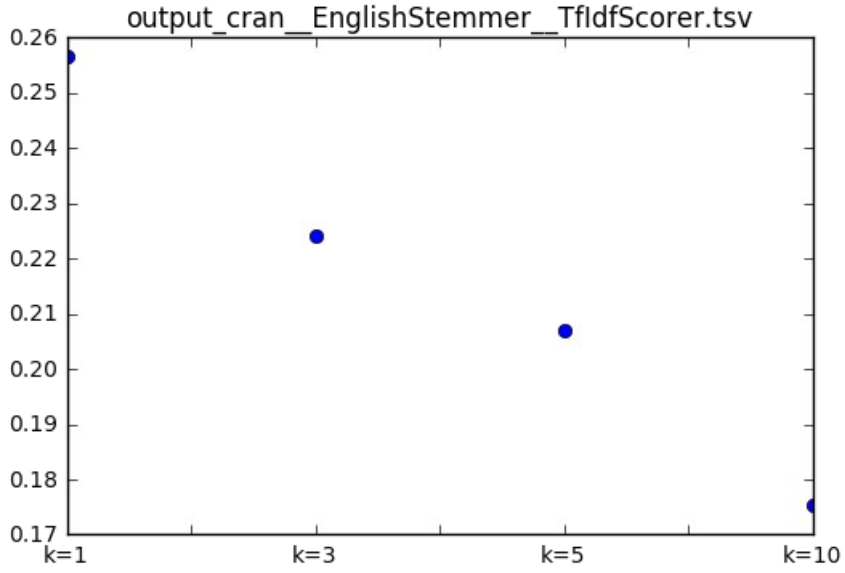

output_cran__EnglishStemmer_stopwords__TfIdfScorer.tsv
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
k=1
k=3
k=5
k=10


output_cran__defaultStemmer__TfIdfscorer.tsv
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
k=1
k=3
k=5
k=10


output_cran__EnglishStemmer_stopwords__CountScorer.tsv
0.26
0.24
0.22
0.20
0.18
0.16
0.14
k=1
k=3
k=5
k=10


output_cran__EnglishStemmer_stopwords__BM25Scorer.tsv
0.32
0.31
0.30
0.29
0.28
0.27
0.26
0.25
0.24
k=1
k=3
k=5
k=10


output_cran__EnglishStemmer__BM25Scorer.tsv
0.32
0.30
0.28
0.26
0.24
0.22
k=1
k=3
k=5
k=10


output_cran__EnglishStemmer__TfIdfScorer.tsv
0.26
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
k=1
k=3
k=5
k=10

9. An answer to each of the following questions:

Which is the best combination stemmer-scorer_function?
English Stemmer-BM25 Scorer (highest Avg-R Precision and also Ndmcg Score for all values of k)

Which is the best stemmer?
English Stemmer able to filter Stopwords (highest Avg-R precision score)
Which is the best scorer function?
BM25 Scorer (highest Avg-R Precision score for all Stemmer configuration)