DMT HW I

Manfredi Roesler Franz - Vlado Vukovic

4/19/2017

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

Number of queries is 225 while number of documents is 1400.

2. A list of used stemmers.

- Default
- Default English
- English with stop words

3. A list of used scorer functions.

- CountScorer
- TfIdfScorer
- BM25Scorer

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 HWdefStem.collection

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

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

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

java it.unimi.di.big.mg4j.tool.IndexBuilder –downcase -S HWdefStem.collection HWdefStem

java it.unimi.di.big.mg4j.tool.IndexBuilder -t it.unimi.di.big.mg4j.index.snowball.EnglishStemmer -S HWengStem.collection HWengStem

java it.unimi.di.big.mg4j.tool.IndexBuilder -t homework.EnglishStemmerStopwords -S HWengStemStop.collection HWengStemStop

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

java homework.RunAllQueries_HW HWdefStem ./Cranfield_Dataset/ais.collection.all_queries.tsv CountScorer text_and_title CountDefStem.tsv

java homework.RunAllQueries_HW HWdefStem ./Cranfield_Dataset/ais.collection.all_queries.tsv TfIdfScorer text_and_title TFIDFDefStem.tsv

java homework.RunAllQueries_HW HWdefStem ./Cranfield_Dataset/ais.collection.all_queries.tsv BM25Scorer text_and_title BM25DefStem.tsv

java homework.RunAllQueries_HW HWengStem ./Cranfield_Dataset/ais.collection.all_queries.tsv CountScorer text_and_title CountEngStem.tsv

java homework.RunAllQueries_HW HWengStem ./Cranfield_Dataset/ais.collection.all_queries.tsv TfIdfScorer text_and_title TFIDFEngStem.tsv

java homework.RunAllQueries_HW HWengStem ./Cranfield_Dataset/ais.collection.all_queries.tsv BM25Scorer text_and_title BM25EngStem.tsv

java homework.RunAllQueries_HW HWengStemStop ./Cranfield_Dataset/ais.collection.all_queries.tsv CountScorer text_and_title CountEngStemStop.tsv

java homework.RunAllQueries_HW HWengStemStop ./Cranfield_Dataset/ais.collection.all_queries.tsv TfIdfScorer text_and_title TFIDFEngStemStop.tsv

java homework.RunAllQueries_HW HWengStemStop ./Cranfield_Dataset/ais.collection.all_queries.tsv BM25Scorer text_and_title BM25EngStemStop.tsv

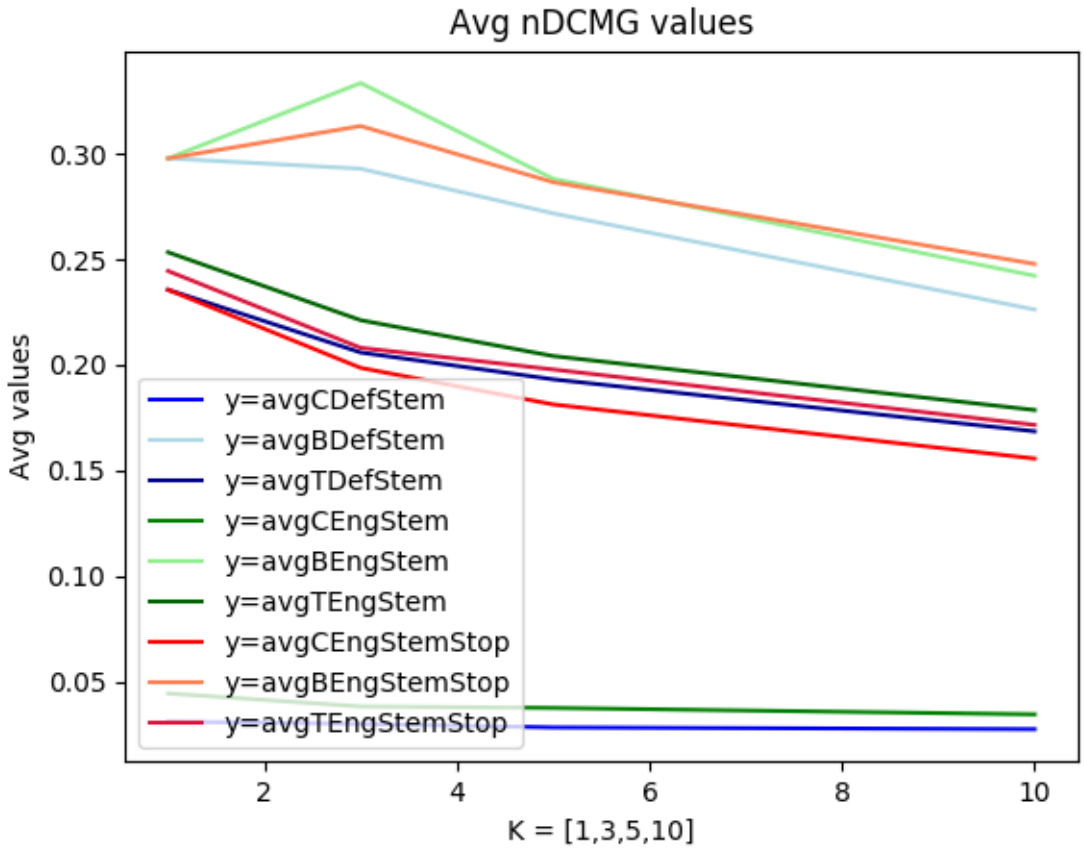
java homework.RunAllQueries_HW HWengStemStop ./Cranfield_Dataset/ais.collection.all_queries.tsv BM25Scorer text aggText.tsv

java homework.RunAllQueries_HW HWengStemStop ./Cranfield_Dataset/ais.collection.all_queries.tsv BM25Scorer title aggTitle.tsv

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

- 0.33585843105383345
- 0.79623040050626237
- 0.72995503134583573
- 0.37778829897220717
- 0.81879606217537237
- 0.77594795455715004
- 0.77909961303064768
- 0.83457707384144164
- 0.80389513849283945
- 0.24544129561370939

8. The plot of the average nMDCG that have: 1) on the 'x' axis the value of k. Consider only the following values: 1, 3, 5, 10. 2) on the 'y' axis the average value of nMDCG: average value over all queries. 3) one curve for each stemmer-scorer_function configuration: 3X3=9 curves in total.



Figure

9. An answer to each of the following questions:

Which is the best combination stemmer-scorer_function? Default English stemmer and BM25Scorer scorer although Default English stemmer with stopwords and BM25Scorer scorer have same result except for topK = 3.

Which is the best stemmer? English default stemmer although have same result except for topK = 3.

Which is the best scorer function? BM25Scorer scorer function is the best with every stemmer configuration.