Homework 1

Ilario Francesco 1469228
Marzio Monticelli 1459333

April 19, 2017

1 Dataset, Stemmers, and Scorer functions

The given Cranfield data-set contains 1400 documents, with a total weight of 1, 5 MB, and we have to evaluate 222 queries.
The stemmers we have to use are the following:

- Default Stemmer;
- English Stemmer;
- English Stemmer able to filter Stopwords.

The scorer functions we have to use are the following:

- Count Scorer;
- BM25 Scorer;
- TfIdf Scorer.

2 Software

2.1 Additional Dependencies

- jfreechart: 1.0.0
- jcommon: 1.0.0
- Apache common-CSV: 1.4

2.2 Scripts

2.2.1 Create the collection

This first script, named `1-createCollection.sh`, will create the directory `Collection` and the collection file.

```
1 DIR="./"
2 DIR_COLL=$DIR"Collection/"
3
4 mkdir $DIR_COLL
5
6 find cran -iname \*.html | java it.unimi.di.big.mg4j.document.FileSetDocumentCollection -f
      HtmlDocumentFactory -p encoding=UTF-8 $DIR_COLL"cran.collection"
```

2.2.2 Create the Inverted Indexes

The script `2-createIndexes.sh` creates all the different indexes that we need. It will put them in different directories, every one named after the used Stemmer.

```
DIR="./"
DIR_COLL=$DIR"Collection/"

DIR_INDX=$DIR"Indexes/"
DIR_INDX_DS=$DIR_INDX"DefautStemmer/"
DIR_INDX_ES=$DIR_INDX"EnglishStemmer/"
DIR_INDX_ESS=$DIR_INDX"EnglishStemmerStopwords/"

mkdir $DIR_INDX
mkdir $DIR_INDX_DS
mkdir $DIR_INDX_ES
mkdir $DIR_INDX_ESS

cp $DIR_COLL"cran.collection" $DIR_INDX_DS"cran.collection"
cp $DIR_COLL"cran.collection" $DIR_INDX_ES"cran.collection"
cp $DIR_COLL"cran.collection" $DIR_INDX_ESS"cran.collection"

java it.unimi.di.big.mg4j.tool.IndexBuilder -S $DIR_COLL"cran.collection" $DIR_INDX_DS"cran"

java it.unimi.di.big.mg4j.tool.IndexBuilder -t
     it.unimi.di.big.mg4j.index.snowball.EnglishStemmer -S $DIR_COLL"cran.collection"
     $DIR_INDX_ES"cran"

java it.unimi.di.big.mg4j.tool.IndexBuilder -t homework.EnglishStemmerStopwords -S
     $DIR_COLL"cran.collection" $DIR_INDX_ESS"cran"
```

2.2.3 Obtain the results

In order to obtain the results, we need to execute different scripts. The first is `3-obtainResults.sh`. It evaluates all the given queries -using the RunAllQueries_HW software- and saves the results in an ad-hoc output file in an ad-hoc directory. When the script finishes, we find a tree of directories, one for each stemmer, containing three files, one for each "field" between "text", "title", and "text_and_title".

```
DIR="./"
DIR_RES=$DIR"QueriesResults/"
DIR_COLL=$DIR"Collection/"

DIR_INDX=$DIR"Indexes/"
DIR_INDX_DS=$DIR_INDX"DefautStemmer/"
DIR_INDX_ES=$DIR_INDX"EnglishStemmer/"
DIR_INDX_ESS=$DIR_INDX"EnglishStemmerStopwords/"

mkdir $DIR_RES

################## Default Stemmer Stopwords ##########################
DIR_DS=$DIR_RES"DefaultStemmer/"
DIR_DS_Te=$DIR_DS"DSText/"
DIR_DS_Ti=$DIR_DS"DSTitle/"
DIR_DS_TT=$DIR_DS"DSTextAndTitle/"

mkdir $DIR_DS
mkdir $DIR_DS_Te
mkdir $DIR_DS_Ti
mkdir $DIR_DS_TT

# text
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
     "CountScorer" "text" $DIR_DS_Te"output_text_DefaultStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
     "TfIdfScorer" "text" $DIR_DS_Te"output_text_DefaultStemmer_TfIdfScorer.tsv"
```

```
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "text" $DIR_DS_Te"output_text_DefaultStemmer_BM25Scorer.tsv"

# title
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "title" $DIR_DS_Ti"output_title_DefaultStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "TfIdfScorer" "title" $DIR_DS_Ti"output_title_DefaultStemmer_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "title" $DIR_DS_Ti"output_title_DefaultStemmer_BM25Scorer.tsv"

# title and text
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "text_and_title"
      $DIR_DS_TT"output_TextAndTitle_DefaultStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "TfIdfScorer" "text_and_title"
      $DIR_DS_TT"output_TextAndTitle_DefaultStemmer_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_DS"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "text_and_title"
      $DIR_DS_TT"output_TextAndTitle_DefaultStemmer_BM25Scorer.tsv"

################## English Stemmer ##############################
DIR_ES=$DIR_RES"EnglishStemmer/"
DIR_ES_Te=$DIR_ES"ESText/"
DIR_ES_Ti=$DIR_ES"ESTitle/"
DIR_ES_TT=$DIR_ES"ESTextAndTitle/"

mkdir $DIR_ES
mkdir $DIR_ES_Te
mkdir $DIR_ES_Ti
mkdir $DIR_ES_TT

# text
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "text" $DIR_ES_Te"output_text_EnglishStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "TfIdfScorer" "text" $DIR_ES_Te"output_text_EnglishStemmer_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "text" $DIR_ES_Te"output_text_EnglishStemmer_BM25Scorer.tsv"

# title
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "title" $DIR_ES_Ti"output_title_EnglishStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "TfIdfScorer" "title" $DIR_ES_Ti"output_title_EnglishStemmer_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "title" $DIR_ES_Ti"output_title_EnglishStemmer_BM25Scorer.tsv"

# title and text
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "text_and_title"
      $DIR_ES_TT"output_TextAndTitle_EnglishStemmer_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "TfIdfScorer" "text_and_title"
      $DIR_ES_TT"output_TextAndTitle_EnglishStemmer_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ES"cran" $DIR"Queries/cran_all_queries.tsv"
      "BM25Scorer" "text_and_title"
      $DIR_ES_TT"output_TextAndTitle_EnglishStemmer_BM25Scorer.tsv"

################### English Stemmer Stopwords ############################
DIR_ESS=$DIR_RES"EnglishStemmerStopwords/"
DIR_ESS_Te=$DIR_ESS"ESSText/"
DIR_ESS_Ti=$DIR_ESS"ESSTitle/"
DIR_ESS_TT=$DIR_ESS"ESSTextAndTitle/"

mkdir $DIR_ESS
mkdir $DIR_ESS_Te
mkdir $DIR_ESS_Ti
mkdir $DIR_ESS_TT

# text
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
      "CountScorer" "text" $DIR_ESS_Te"output_text_EnglishStemmerStopwords_CountScorer.tsv"
```

```
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "TfIdfScorer" "text" $DIR_ESS_Te"output_text_EnglishStemmerStopwords_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "BM25Scorer" "text" $DIR_ESS_Te"output_text_EnglishStemmerStopwords_BM25Scorer.tsv"

# title
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "CountScorer" "title" $DIR_ESS_Ti"output_title_EnglishStemmerStopwords_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "TfIdfScorer" "title" $DIR_ESS_Ti"output_title_EnglishStemmerStopwords_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "BM25Scorer" "title" $DIR_ESS_Ti"output_title_EnglishStemmerStopwords_BM25Scorer.tsv"

# title and text
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "CountScorer" "text_and_title"
     $DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_CountScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "TfIdfScorer" "text_and_title"
     $DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_TfIdfScorer.tsv"
java homework.RunAllQueries_HW $DIR_INDX_ESS"cran" $DIR"Queries/cran_all_queries.tsv"
     "BM25Scorer" "text_and_title"
     $DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_BM25Scorer.tsv"
```

Subsequently, we can run the script `4-performScoring.sh`. This script calculates the Average R-Precision and the nMDCG on results obtained with the field "text_and_title" for every Stemmer-Scorer function pairs.

```
### scoring ###
if [[ $# -eq 1 ]]; then
   MD=$1
elif [[ $# -eq 2 ]]; then
   MD=$1
   GTRU=$2
else
   MD="./QueriesResults"
   GTRU="./Queries/cran_Ground_Truth.tsv"
fi


bold=$(tput bold)
normal=$(tput sgr0)

K=(1 3 5 10);

echo ${bold}"################ AverageRPRecision ################"
echo "Default Stemmer - BM25Scorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/DefaultStemmer/DSTextAndTitle/
     output_TextAndTitle_DefaultStemmer_BM25Scorer.tsv" $GTRU

echo
echo ${bold}"English Stemmer - BM25Scorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmer/ESTextAndTitle/
     output_TextAndTitle_EnglishStemmer_BM25Scorer.tsv" $GTRU

echo
echo ${bold}"English Stopwords Stemmer - BM25Scorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmerStopwords/ESSTextAndTitle/
     output_TextAndTitle_EnglishStemmerStopwords_BM25Scorer.tsv" $GTRU

echo
echo ${bold}"Default Stemmer - CountScorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/DefaultStemmer/DSTextAndTitle/
     output_TextAndTitle_DefaultStemmer_CountScorer.tsv" $GTRU

echo
echo ${bold}"English Stemmer - CountScorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmer/ESTextAndTitle/
     output_TextAndTitle_EnglishStemmer_CountScorer.tsv" $GTRU

echo
echo ${bold}"English Stopwords Stemmer - CountScorer"${normal}
```

```
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmerStopwords/ESSTextAndTitle/output_TextAndTitle_EnglishStemmerStopwords_CountScorer.tsv" $GTRU

echo
echo ${bold}"Default Stemmer - TfIdfScorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/DefaultStemmer/DSTextAndTitle/output_TextAndTitle_DefaultStemmer_TfIdfScorer.tsv" $GTRU

echo
echo ${bold}"English Stemmer - TfIdfScorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmer/ESTextAndTitle/output_TextAndTitle_EnglishStemmer_TfIdfScorer.tsv" $GTRU

echo
echo ${bold}"English Stopwords Stemmer - TfIdfScorer"${normal}
java com.weird.hw1.ese.AverageRPrecision $MD"/EnglishStemmerStopwords/ESSTextAndTitle/output_TextAndTitle_EnglishStemmerStopwords_TfIdfScorer.tsv" $GTRU

echo
echo ${bold}"################ nMDCG ################"
echo "Default Stemmer - BM25Scorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/DefaultStemmer/DSTextAndTitle/output_TextAndTitle_DefaultStemmer_BM25Scorer.tsv" $GTRU
done

echo
echo ${bold}"English Stemmer - BM25Scorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmer/ESTextAndTitle/output_TextAndTitle_EnglishStemmer_BM25Scorer.tsv" $GTRU
done

echo
echo ${bold}"English Stopwords Stemmer - BM25Scorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmerStopwords/ESSTextAndTitle/output_TextAndTitle_EnglishStemmerStopwords_BM25Scorer.tsv" $GTRU
done

echo
echo ${bold}"Default Stemmer - CountScorer"${normal}
for k in ${K[*]}; do
java com.weird.hw1.ese.nMDCG -k $k $MD"/DefaultStemmer/DSTextAndTitle/output_TextAndTitle_DefaultStemmer_CountScorer.tsv" $GTRU
done

echo
echo ${bold}"English Stemmer - CountScorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmer/ESTextAndTitle/output_TextAndTitle_EnglishStemmer_CountScorer.tsv" $GTRU
done

echo
echo ${bold}"English Stopwords Stemmer - CountScorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmerStopwords/ESSTextAndTitle/output_TextAndTitle_EnglishStemmerStopwords_CountScorer.tsv" $GTRU
done

echo
echo ${bold}"Default Stemmer - TfIdfScorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/DefaultStemmer/DSTextAndTitle/output_TextAndTitle_DefaultStemmer_TfIdfScorer.tsv" $GTRU
done

echo
echo ${bold}"English Stemmer - TfIdfScorer"${normal}
for k in ${K[*]}; do
   java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmer/ESTextAndTitle/output_TextAndTitle_EnglishStemmer_TfIdfScorer.tsv" $GTRU
done
```

```
102
103 echo
104 echo ${bold}"English Stopwords Stemmer - TfIdfScorer"${normal}
105 for k in ${K[*]}; do
106    java com.weird.hw1.ese.nMDCG -k $k $MD"/EnglishStemmerStopwords/ESSTextAndTitle/
          output_TextAndTitle_EnglishStemmerStopwords_TfIdfScorer.tsv" $GTRU
107 done
```

The third script -5-ranksAggregation.sh- makes the aggregation of the queries results on separate fields Text, and Title. It uses by default an adaptive value of K (it will be maximum possible for each query), the English Stemmer with Stopwords, and BM25 Scorer function.

```
1  DIR="./"
2  DIR_ESS=$DIR"QueriesResults/EnglishStemmerStopwords/"
3  DIR_ESS_Te=$DIR_ESS"ESSText/"
4  DIR_ESS_Ti=$DIR_ESS"ESSTitle/"
5
6  F_ESS_Te=$DIR_ESS_Te"output_text_EnglishStemmerStopwords_BM25Scorer.tsv"
7  F_ESS_Ti=$DIR_ESS_Ti"output_title_EnglishStemmerStopwords_BM25Scorer.tsv"
8  F_ESS_A=$DIR_ESS"output_aggregated_EnglishStemmerStopwords_BM25Scorer.tsv"
9
10 GTRU="./Queries/cran_Ground_Truth.tsv"
11 R="-r 2"
12 K=""
13 V=""
14
15 while [[ $# -gt 1 ]]
16 do
17    key="$1"
18
19    case $key in
20       -k|-K)
21       K="-k "$2
22       shift
23       ;;
24       -R|-r|--ratio)
25       R="-r "$2
26       shift
27       ;;
28       -t|--text)
29       F_ESS_Te=$2
30       shift
31       ;;
32       -T|--title)
33       F_ESS_Ti=$2
34       shift
35       ;;
36       -o|--output)
37       F_ESS_A=$2
38       shift
39       ;;
40       -v|--verbose)
41          V=--verbose
42          ;;
43       -V|--Vverbose)
44          V=--Vverbose
45          ;;
46       *)
47       "usage: "$BASH_SOURCE" [-k|-K K_VALUE -R|--ratio RATIO -t|--text Text_score_FILE
               -T|--title Title_score_FILE -o|--output Output_File]"
48       exit 1
49       ;;
50    esac
51    shift
52 done
53
54 echo
55 # Fagin's Algorithm
56 java com.weird.hw1.rank_aggregation.FaginsAlgorithm $K $R $F_ESS_Te $F_ESS_Ti $F_ESS_AVAG $V
57
```

```
echo -ne $bold"Fagin's Algorithm"$normal" "
java com.weird.hw1.ese.AverageRPrecision $F_ESS_AVAG $GTRU

# Fagin's Threshold Algorithm
java com.weird.hw1.rank_aggregation.ThresholdAlgorithm2 $K $R $F_ESS_Te $F_ESS_Ti
     $F_ESS_ATHR $V

echo -ne $bold"Fagin's Threshold Algorithm"$normal" "
java com.weird.hw1.ese.AverageRPrecision $F_ESS_ATHR $GTRU
```

At the end, it is possible to plot the nMDCG results using the following script (`6-plot.sh`):

```
# Constants
DIROUT="./Plots/"
mkdir $DIROUT

DIR_M="./"
DIR_QRES=$DIR_M"QueriesResults/"
DIR_RES=$DIR_M"Queries/"

DIR_DS_TT=$DIR_QRES"DefaultStemmer/DSTextAndTitle/"
DIR_ES_TT=$DIR_QRES"EnglishStemmer/ESTextAndTitle/"
DIR_ES_TT=$DIR_QRES"EnglishStemmerStopwords/ESSTextAndTitle/"
DSBM25=$DIR_DS_TT"output_TextAndTitle_DefaultStemmer_BM25Scorer.tsv"
DSCOUN=$DIR_DS_TT"output_TextAndTitle_DefaultStemmer_CountScorer.tsv"
DSTFIF=$DIR_DS_TT"output_TextAndTitle_DefaultStemmer_TfIdfScorer.tsv"

ESBM25=$DIR_ES_TT"output_TextAndTitle_EnglishStemmer_BM25Scorer.tsv"
ESCOUN=$DIR_ES_TT"output_TextAndTitle_EnglishStemmer_CountScorer.tsv"
ESTFIF=$DIR_ES_TT"output_TextAndTitle_EnglishStemmer_TfIdfScorer.tsv"

ESSBM25=$DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_BM25Scorer.tsv"
ESSCOUN=$DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_CountScorer.tsv"
ESSTFIF=$DIR_ESS_TT"output_TextAndTitle_EnglishStemmerStopwords_TfIdfScorer.tsv"

GTRU=$DIR_RES"cran_Ground_Truth.tsv"


K="1 3 5 10"

bold=$(tput bold)
normal=$(tput sgr0)

# Bar Chart
echo
echo $bold"########## BarChart ##########"$normal
java com.weird.hw1.plot.BarPlotHWnMDCGResults -K $K -f "$DSBM25" "Default Stemmer" "BM25"
     "$DSCOUN" "Default Stemmer" "Count" "$DSTFIF" "Default Stemmer" "TfIdf" "$ESBM25"
     "English Stemmer" "BM25" "$ESCOUN" "English Stemmer" "Count" "$ESTFIF" "English
     Stemmer" "TfIdf" "$ESSBM25" "English Stopwords Stemmer" "BM25" "$ESSCOUN" "English
     Stopwords Stemmer" "Count" "$ESSTFIF" "English Stopwords Stemmer" "TfIdf" -g "$GTRU"
     -o $DIROUT"barChart_All.jpeg"

# Curves
# Default Stemmer
echo $bold"########## Curves ##########"$normal
echo
echo $bold"########## Default Stemmer ##########"$normal
java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$DSBM25" -g "$GTRU" -o
     $DIROUT"plot_DefaultStemmer_BM25Scorer.jpeg" -T "Default Stemmer - BM25" -c "blue"
echo
java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$DSCOUN" -g "$GTRU" -o
     $DIROUT"plot_DefaultStemmer_CountScorer.jpeg" -T "Default Stemmer - Count" -c "red"
echo
java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$DSTFIF" -g "$GTRU" -o
     $DIROUT"plot_DefaultStemmer_TfIdfScorer.jpeg" -T "Default Stemmer - TfIdf" -c "green"

# English Stemmer
echo
echo $bold"########## English Stemmer ##########"$normal
java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESBM25" -g "$GTRU" -o
```

```
      $DIROUT"plot_EnglishStemmer_BM25Scorer.jpeg" -T "English Stemmer - BM25" -c "blue"
52 echo
53 java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESCOUN" -g "$GTRU" -o
      $DIROUT"plot_EnglishStemmer_CountScorer.jpeg" -T "English Stemmer - Count" -c "red"
54 echo
55 java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESTFIF" -g "$GTRU" -o
      $DIROUT"plot_EnglishStemmer_TfIdfScorer.jpeg" -T "English Stemmer - TfIdf" -c "green"
56
57 # English Stemmer Stopwords
58 echo
59 echo $bold"########## English Stemmer with Stopwords ##########"$normal
60 java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESSBM25" -g "$GTRU" -o
      $DIROUT"plot_EnglishStemmerStopwords_BM25Scorer.jpeg" -T "English Stemmer with
      Stopwords - BM25" -c "blue"
61 echo
62 java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESSCOUN" -g "$GTRU" -o
      $DIROUT"plot_EnglishStemmerStopwords_CountScorer.jpeg" -T "English Stemmer with
      Stopwords - Count" -c "red"
63 echo
64 java com.weird.hw1.plot.CurvePlotternMDCG -K $K -e "$ESSTFIF" -g "$GTRU" -o
      $DIROUT"plot_EnglishStemmerStopwords_TfIdfScorer.jpeg" -T "English Stemmer with
      Stopwords - TfIdf" -c "green"
```

2.3 Little Tip

We suggest to use the script `0-executeAll.sh` in order to run all the previous ones. It asks to you to tune some parameters or use the default values.

```
1  K_DEF=5
2  re='^[0-9]+$'
3  V=""
4
5  NUM_SCRIPTS=6
6
7
8  bold=$(tput bold)
9  normal=$(tput sgr0)
10
11 # Preparsing Help
12 if [[ $# -ge 1 ]]; then
13    for arg in $@; do
14       if [[ $arg == "--help" || $arg == "-h" ]]; then
15          me=$BASH_SOURCE
16          echo "$ source" $me":"
17          echo "  -h --help     prints this help"
18          echo "  -v --verbose  sets the verbose mode"
19       fi
20    done
21 fi
22
23 # Handle verbose mode
24 if [[ $# -ne 1 || ( $1 != "-v" && $1 != "--verbose" ) ]]; then
25    V=" > /dev/null"
26    echo ${bold}"[1/"$NUM_SCRIPTS"]"${normal}" Performing collection creation"
27 fi
28
29 eval source 1-createCollection.sh $V
30
31 if [[ $V != "" ]]; then
32    echo ${bold}"[2/"$NUM_SCRIPTS"]"${normal}" Performing Indexes creation"
33 fi
34 eval source 2-createIndexes.sh $V
35
36 if [[ $V != "" ]]; then
37    echo ${bold}"[3/"$NUM_SCRIPTS"]"${normal}" Obtaining Results"
38 fi
39
40 eval source 3-obtainResults.sh $V
41
42 echo
```

```
43 echo ${bold}"[4/"$NUM_SCRIPTS"]"${normal}" Press Enter to perform scoring"
44 read
45 clear
46
47 source 4-performScoring.sh
48
49 # grub K Value
50 echo
51 echo ${bold}"[5/"$NUM_SCRIPTS"]"${normal}" Rank Aggreagation"
52 while (true); do
53    echo -n "which value for K do you want to use? [Press ENTER to use the max possible K
          value for each instance] "
54    read K
55    if [[ $K =~ $re ]]; then
56       K=-k $K
57       break;
58    elif [[ $K == "" ]]; then
59       K=""
60       break;
61    else
62       echo "[ERR] \""$K"\" is not a number. Retry" >&2;
63    fi
64 done
65
66 # grub Ratio value
67 while (true); do
68    echo -n "which value for Ratio (importance of Title wrt Text) do you want to use?
          [Default=2] "
69    read R
70    if [[ $R =~ $re ]]; then
71       R=-r $R
72       break;
73    elif [[ $R == "" ]]; then
74       R=""
75       break;
76    else
77       echo "[ERR] \""$R"\" is not a number. Retry" >&2;
78    fi
79 done
80
81 eval source 5-ranksAggregation.sh $K $R
82
83 echo
84 echo ${bold}"[6/"$NUM_SCRIPTS"]"${normal}" Plot Results"
85 eval source 6-plot.sh
```

3 Results

3.1 Average R-Precision

	Default	English	English With Stopwords
BM25	0.255	0.262	0.266
Count	0.023	0.030	0.160
TfIdf	0.179	0.189	0.191

Table 1: Average R-Precision for Scorers

	English With Stopwords - BM25
Fagin	0.249
Threshold	0.249

Table 2: Average R-Precision for Ranks Aggregations

3.2 More Rank Aggregations Results

The following results are obtained using the English Stemmer with Stopwords and the BM25 Scorer Function.

	1	3	5	10	20	∞
Fagin	0.067	0.178	0.215	0.243	0.248	0.249
Threshold	0.067	0.178	0.215	0.243	0.248	0.249

Table 3: More Average R-Precision for Ranks Aggregations

4 Plots

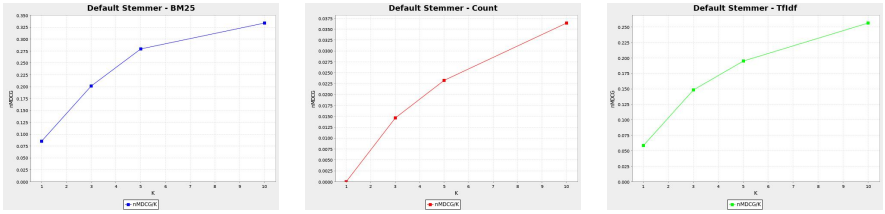


Figure 1: Default Stemmer

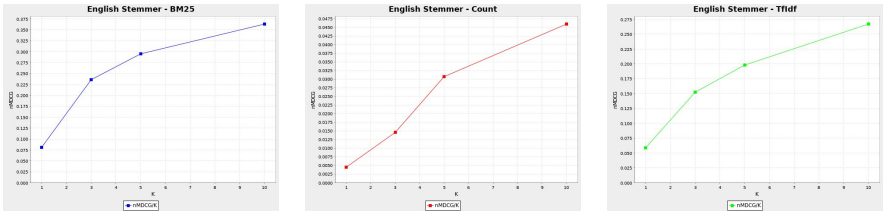


Figure 2: English Stemmer

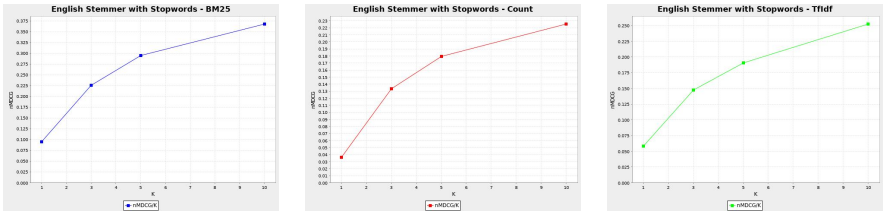
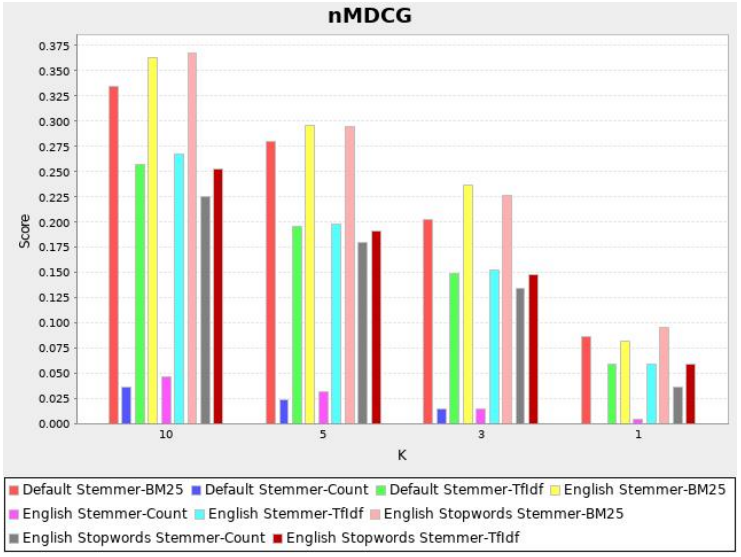


Figure 3: English Stemmer Stopwords



5 Answers

5.1 Best Stemmer-Scorer Function combination

	Default	English	English With Stopwords
BM25	0.255	0.262	0.266
Count	0.022	0.028	0.159
TfIdf	0.179	0.189	0.191

Table 4: (Scorer Functions / Stemmer) Using Average RPrecision

In order to extract this data the K value has been set to 0. Doing this, the software will always use the greatest K possible.

	Default	English	English With Stopwords
BM25	0.406	0.438	0.443
Count	0.058	0.069	0.295
TfIdf	0.318	0.331	0.326

Table 5: (Scorer Functions / Stemmer) Using nMDCG

From tables 4 and 5, we can notice that the best Stemmer-Scorer Function combination is **English Stemmer with Stopwords** and **BM25**.

5.2 Best Stemmer

	Default	English	English With Stopwords
BM25	0.255	0.262	0.266
Count	0.022	0.028	0.159
TfIdf	0.179	0.189	0.191

Table 6: (Scorer Functions / Stemmer) Using Average RPrecision

In order to extract this data the K value has been set to 0. Doing this, the software will always use the greatest K possible.

	Default	English	English With Stopwords
BM25	0.406	0.438	0.443
Count	0.058	0.069	0.295
TfIdf	0.318	0.331	0.326

Table 7: (Scorer Functions / Stemmer) Using nMDCG

From tables 6 and 7, we can notice that the best stemmer is **English Stemmer with Stopwords**.

5.3 Best Scorer Function

	Default	English	English With Stopwords
BM25	0.255	0.262	0.266
Count	0.022	0.028	0.159
TfIdf	0.179	0.189	0.191

Table 8: (Scorer Functions / Stemmer) Using Average RPrecision

In order to extract this data the K value has been set to 0. Doing this, the software will always use the greatest K possible.

	Default	English	English With Stopwords
BM25	0.406	0.438	0.443
Count	0.058	0.069	0.295
TfIdf	0.318	0.331	0.326

Table 9: (Scorer Functions / Stemmer) Using nMDCG

From tables 8 and 9, we can notice that the best Scorer Function is **BM25**.