Information Retrieval – Lucene and Trec_Eval
Assignment – 1

Pallavit Aggarwal – 22333721, IS

About Lucene

Lucene's primary goal is to facilitate information retrieval. In order for Lucene to know what "words" are, it analyzes the text during indexing, extracting it to terms. Analysis, in Lucene, is the process of converting field text into its most fundamental indexed representation, terms.

In general, an analyzer tokenizes text by performing any different operations on it, which could include extracting words, discarding punctuation, removing accents from characters, lowercasing (normalizing), removing common words, reducing words to a root form (stemming), or changing words into the basic form (lemmatization). This process is also called tokenization, and the chunks of text pulled from a stream of text are called tokens. Tokens, combined with their associated field name, are terms.

Implementation

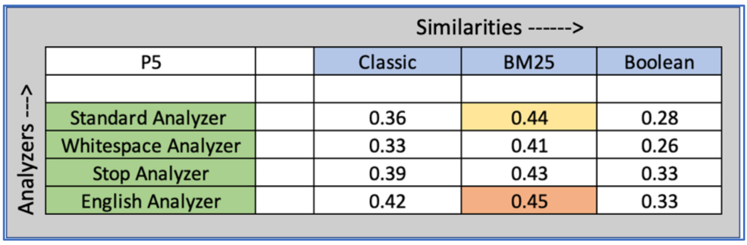
Our Scenario is Ad-hoc IR which assumes fixed document collection over a dynamic one, and where the queries cannot be refined.

We use Java 7 as the choice of language of implementation. Plugin and Dependency management is done using Maven, and the libraries used are: lucene-core:8.6.3, lucene-queryparser:8.6.3 and lucene-analyzers-common:8.10.0.

We first read the file cran.all.1400 . This file has information split .I <number> where number is 1 – 1400 and values like .I (index) , .T (title), .A (author), .W (words) and .B (bibliography). Then we add documents containing Fields to IndexWriter which analyzes the documents using the Analyzer and then creates/open/edit indexes as required and stores/updates them in a Directory.

We use **TextField** instead of StringField to have Indexes analyzed. (Index.Field.Analyzed).

P5

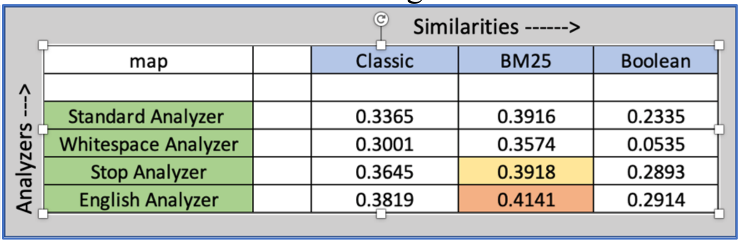


Similarities ------>

Analyzers --->

P5		Classic	BM25	Boolean
Standard Analyzer		0.36	0.44	0.28
Whitespace Analyzer		0.33	0.41	0.26
Stop Analyzer		0.39	0.43	0.33
English Analyzer		0.42	0.45	0.33

Mean Average Precision

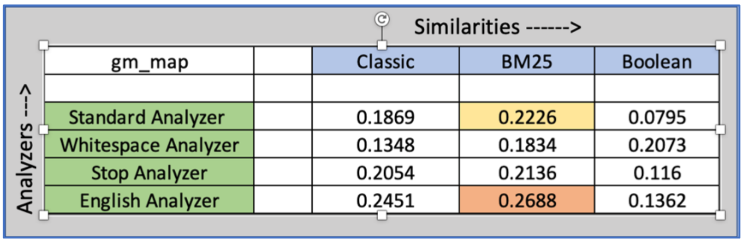


Similarities ------>

Analyzers --->

map		Classic	BM25	Boolean
Standard Analyzer		0.3365	0.3916	0.2335
Whitespace Analyzer		0.3001	0.3574	0.0535
Stop Analyzer		0.3645	0.3918	0.2893
English Analyzer		0.3819	0.4141	0.2914

Geometric Mean MAP

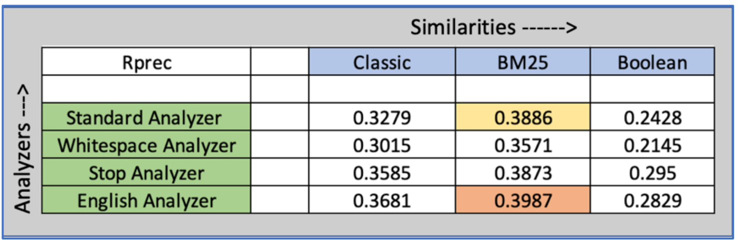


Similarities ------>

Analyzers --->

gm_map		Classic	BM25	Boolean
Standard Analyzer		0.1869	0.2226	0.0795
Whitespace Analyzer		0.1348	0.1834	0.2073
Stop Analyzer		0.2054	0.2136	0.116
English Analyzer		0.2451	0.2688	0.1362

Recall-Precision



Similarities ------>

Analyzers --->

Rprec		Classic	BM25	Boolean
Standard Analyzer		0.3279	0.3886	0.2428
Whitespace Analyzer		0.3015	0.3571	0.2145
Stop Analyzer		0.3585	0.3873	0.295
English Analyzer		0.3681	0.3987	0.2829

Above are the results of different Analyzers (given in green) implemented for different Similarities (given in Blue). The best score and second-best score obtained for each set of values have been highlighted with orange and yellow respectively. **For example**, The P5 score which represents the Precision after 5 docs retrieved has values 0.36 for Standard Analyzer implemented with Classic

Similarity, 0.44 for BM25Similarity and 0.28 for BooleanSimilarity but the maximum P5 score observed was for English Analyzer with BM25Similarity.

Analyzers

- **Whitespace Analyzer** uses whitespace tokenizer only, no stop word or case-sensitive filter is used.
- **Stop analyzer** uses lower case tokenizer and stop-token filter to remove stop words from token streams.
- **Standard Analyzer** has a tokenizer and a standard(normalizes tokens), lower-case and stop token filter.
- And finally, the **English Analyzer** which implements a standard tokenizer, Standard Filter, EnglishPossesive Filter, Lowercase Filter, Stop filter and PortStemFilter.

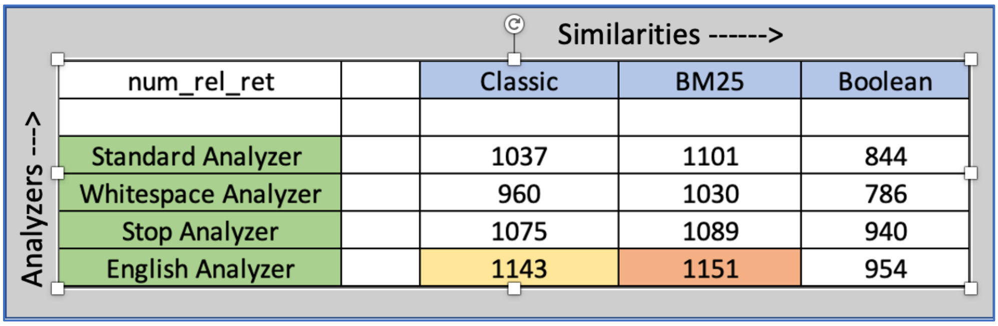
Similarities

Different TF-IDF (Common Locally * Rare Globally) implementations

- **Classic Similarity** is the default similarity which is based on the highly optimized Vector Space Model.
- **BM25 Similarity** is an optimized implementation of the successful Okapi BM25 model (k1=1.2 i.e. term frequency saturation , b=0.75 i.e. document length penalization factor)
- **Boolean Similarity** is a simple similarity that gives terms a score that is equal to their query boost.

Result

The scores above are shown when number of retrieved documents are 11250, and number of queries run are 225. The mean average precision and the recall-precision tell us that English Analyzer performs better than others. This could be due the fact that it implements multiple filters. The P5 score which is the Precision after 5 docs retrieved is also the highest for English Analyzer.



Similarities ------>

Analyzers --->

num_rel_ret		Classic	BM25	Boolean
Standard Analyzer		1037	1101	844
Whitespace Analyzer		960	1030	786
Stop Analyzer		1075	1089	940
English Analyzer		1143	1151	954

Another metric that we can consider is the **num_rel_ret** which gives us the total number of relevant documents retrieved over all queries. Out of 1837 num_rel which were constant throughout all trec_evals, above image shows the number of relevant documents retrieved. Here again, English analyzer with the BM25 default similarity has the highest values and the second highest is also the same analyzer with the classic or default similarity.

The command /.trec_eval ../corpus/QRelsCorrectedforTRECeval.txt ../lucene-db/outputs could've been executed with an -a to provide even more information about the analyzers and their performances like **recall5** and **11-pt_avg**.

As a next step, we can build our own custom analyzer and improve the performance even further for our dataset.

Thank you.

Bibliography

https://www.baeldung.com/lucene-analyzers

https://www.tutorialspoint.com/lucene/lucene_analysis.htm

https://www.infoq.com/articles/similarity-scoring-elasticsearch/

https://www.elastic.co/blog/found-bm-vs-lucene-default-similarity

https://bagua.cct.lsu.edu/dlcurric/modDev/package_modules/MidtermModuleTeam5-TRECevalFinal.pdf

https://stackoverflow.com/questions/2602253/how-does-lucene-index-documents?rq=1

https://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/A.README

https://www.essi.upc.edu/dtim/blog/post/mining-similarity-between-text-documents-using-apache-lucene

https://www.javacodegeeks.com/2015/09/lucene-analysis-process-guide.html

https://github.com/stmunees/Lucene-Information-Retrieval 1/blob/master/src/main/java/IR/Searcher.java

https://ccc.inaoep.mx/~villasen/bib/AN%20OVERVIEW%20OF%20EVALUATION%20METHODS%20IN%20TREC%20AD%20HOC%20IR%20AND%20TREC%20QA.pdf

https://www.lucenetutorial.com/advanced-topics/scoring.html

https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/package-summary.html#scoring

https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

https://nlp.stanford.edu/IR-book/html/htmledition/queries-as-vectors-1.html

https://github.com/olivernn/cranfield

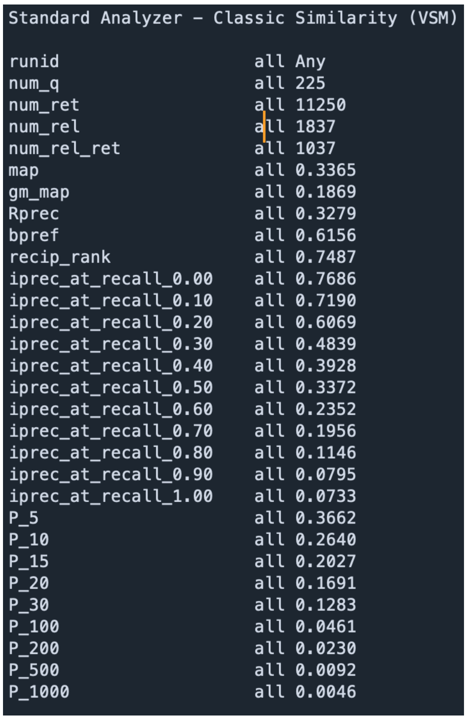
https://github.com/JalajVora/Information-Retrieval-System-using-Apache-Lucene/blob/master/src/LuceneMain.java

https://github.com/ddunne6/lucene-cranfield-collection/blob/master/assignment-1/src/main/java/ie/tcd/ddunne6/QueryIndex.java

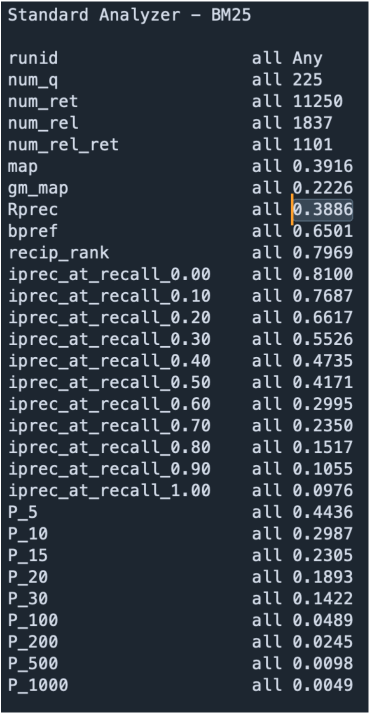
https://github.com/ZhouShengsheng/ci6226-ira-g14

https://github.com/Menotron/cs7is3-cranfield-search/tree/master/src/main/java/ie/tcd/cs7is3/cranfield

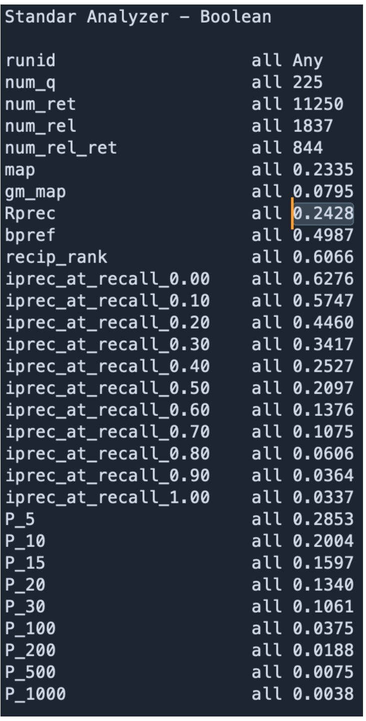
Standard Analyzer scores



```
Standard Analyzer – Classic Similarity (VSM)

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1037
map                     all 0.3365
gm_map                  all 0.1869
Rprec                   all 0.3279
bpref                   all 0.6156
recip_rank              all 0.7487
iprec_at_recall_0.00    all 0.7686
iprec_at_recall_0.10    all 0.7190
iprec_at_recall_0.20    all 0.6069
iprec_at_recall_0.30    all 0.4839
iprec_at_recall_0.40    all 0.3928
iprec_at_recall_0.50    all 0.3372
iprec_at_recall_0.60    all 0.2352
iprec_at_recall_0.70    all 0.1956
iprec_at_recall_0.80    all 0.1146
iprec_at_recall_0.90    all 0.0795
iprec_at_recall_1.00    all 0.0733
P_5                     all 0.3662
P_10                    all 0.2640
P_15                    all 0.2027
P_20                    all 0.1691
P_30                    all 0.1283
P_100                   all 0.0461
P_200                   all 0.0230
P_500                   all 0.0092
P_1000                  all 0.0046
```

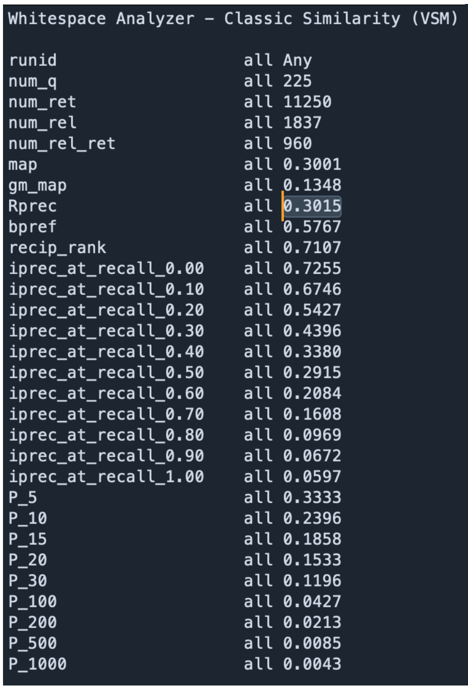


```
Standard Analyzer – BM25

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1101
map                     all 0.3916
gm_map                  all 0.2226
Rprec                   all 0.3886
bpref                   all 0.6501
recip_rank              all 0.7969
iprec_at_recall_0.00    all 0.8100
iprec_at_recall_0.10    all 0.7687
iprec_at_recall_0.20    all 0.6617
iprec_at_recall_0.30    all 0.5526
iprec_at_recall_0.40    all 0.4735
iprec_at_recall_0.50    all 0.4171
iprec_at_recall_0.60    all 0.2995
iprec_at_recall_0.70    all 0.2350
iprec_at_recall_0.80    all 0.1517
iprec_at_recall_0.90    all 0.1055
iprec_at_recall_1.00    all 0.0976
P_5                     all 0.4436
P_10                    all 0.2987
P_15                    all 0.2305
P_20                    all 0.1893
P_30                    all 0.1422
P_100                   all 0.0489
P_200                   all 0.0245
P_500                   all 0.0098
P_1000                  all 0.0049
```

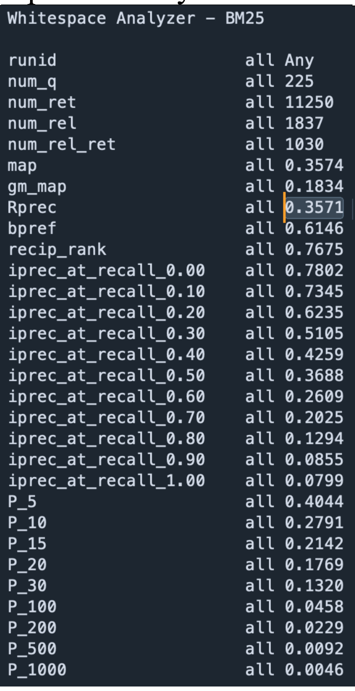


```
Standar Analyzer – Boolean

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 844
map                     all 0.2335
gm_map                  all 0.0795
Rprec                   all 0.2428
bpref                   all 0.4987
recip_rank              all 0.6066
iprec_at_recall_0.00    all 0.6276
iprec_at_recall_0.10    all 0.5747
iprec_at_recall_0.20    all 0.4460
iprec_at_recall_0.30    all 0.3417
iprec_at_recall_0.40    all 0.2527
iprec_at_recall_0.50    all 0.2097
iprec_at_recall_0.60    all 0.1376
iprec_at_recall_0.70    all 0.1075
iprec_at_recall_0.80    all 0.0606
iprec_at_recall_0.90    all 0.0364
iprec_at_recall_1.00    all 0.0337
P_5                     all 0.2853
P_10                    all 0.2004
P_15                    all 0.1597
P_20                    all 0.1340
P_30                    all 0.1061
P_100                   all 0.0375
P_200                   all 0.0188
P_500                   all 0.0075
P_1000                  all 0.0038
```

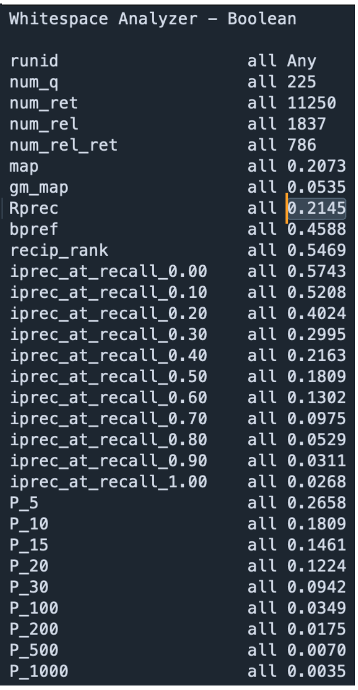
Whitespace Analzyer scores



```
Whitespace Analyzer – Classic Similarity (VSM)

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 960
map                     all 0.3001
gm_map                  all 0.1348
Rprec                   all 0.3015
bpref                   all 0.5767
recip_rank              all 0.7107
iprec_at_recall_0.00    all 0.7255
iprec_at_recall_0.10    all 0.6746
iprec_at_recall_0.20    all 0.5427
iprec_at_recall_0.30    all 0.4396
iprec_at_recall_0.40    all 0.3380
iprec_at_recall_0.50    all 0.2915
iprec_at_recall_0.60    all 0.2084
iprec_at_recall_0.70    all 0.1608
iprec_at_recall_0.80    all 0.0969
iprec_at_recall_0.90    all 0.0672
iprec_at_recall_1.00    all 0.0597
P_5                     all 0.3333
P_10                    all 0.2396
P_15                    all 0.1858
P_20                    all 0.1533
P_30                    all 0.1196
P_100                   all 0.0427
P_200                   all 0.0213
P_500                   all 0.0085
P_1000                  all 0.0043
```

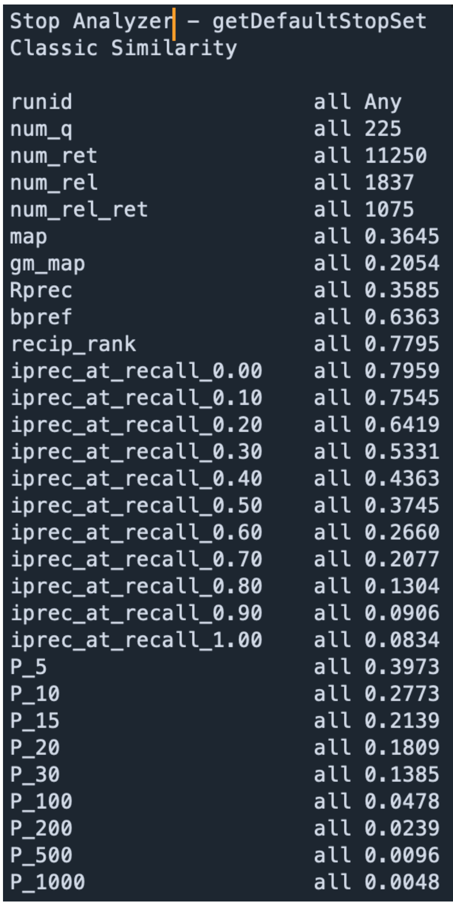


```
Whitespace Analyzer – BM25

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1030
map                     all 0.3574
gm_map                  all 0.1834
Rprec                   all 0.3571
bpref                   all 0.6146
recip_rank              all 0.7675
iprec_at_recall_0.00    all 0.7802
iprec_at_recall_0.10    all 0.7345
iprec_at_recall_0.20    all 0.6235
iprec_at_recall_0.30    all 0.5105
iprec_at_recall_0.40    all 0.4259
iprec_at_recall_0.50    all 0.3688
iprec_at_recall_0.60    all 0.2609
iprec_at_recall_0.70    all 0.2025
iprec_at_recall_0.80    all 0.1294
iprec_at_recall_0.90    all 0.0855
iprec_at_recall_1.00    all 0.0799
P_5                     all 0.4044
P_10                    all 0.2791
P_15                    all 0.2142
P_20                    all 0.1769
P_30                    all 0.1320
P_100                   all 0.0458
P_200                   all 0.0229
P_500                   all 0.0092
P_1000                  all 0.0046
```

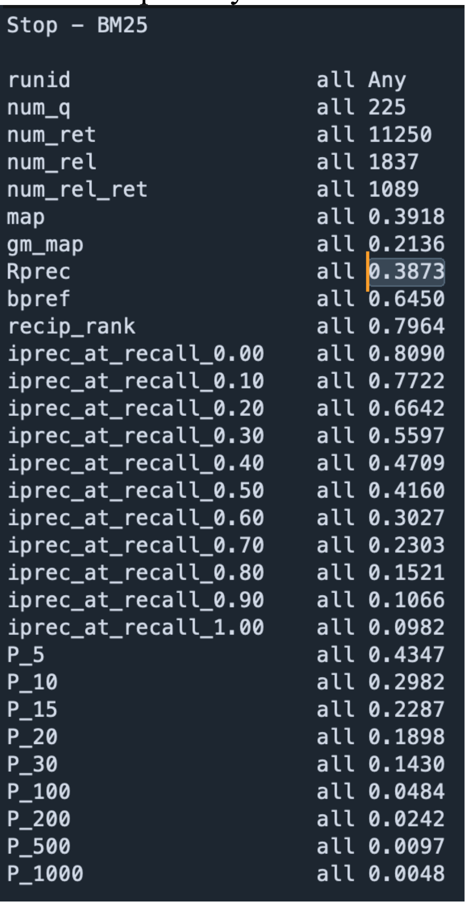


```
Whitespace Analyzer – Boolean

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 786
map                     all 0.2073
gm_map                  all 0.0535
Rprec                   all 0.2145
bpref                   all 0.4588
recip_rank              all 0.5469
iprec_at_recall_0.00    all 0.5743
iprec_at_recall_0.10    all 0.5208
iprec_at_recall_0.20    all 0.4024
iprec_at_recall_0.30    all 0.2995
iprec_at_recall_0.40    all 0.2163
iprec_at_recall_0.50    all 0.1809
iprec_at_recall_0.60    all 0.1302
iprec_at_recall_0.70    all 0.0975
iprec_at_recall_0.80    all 0.0529
iprec_at_recall_0.90    all 0.0311
iprec_at_recall_1.00    all 0.0268
P_5                     all 0.2658
P_10                    all 0.1809
P_15                    all 0.1461
P_20                    all 0.1224
P_30                    all 0.0942
P_100                   all 0.0349
P_200                   all 0.0175
P_500                   all 0.0070
P_1000                  all 0.0035
```

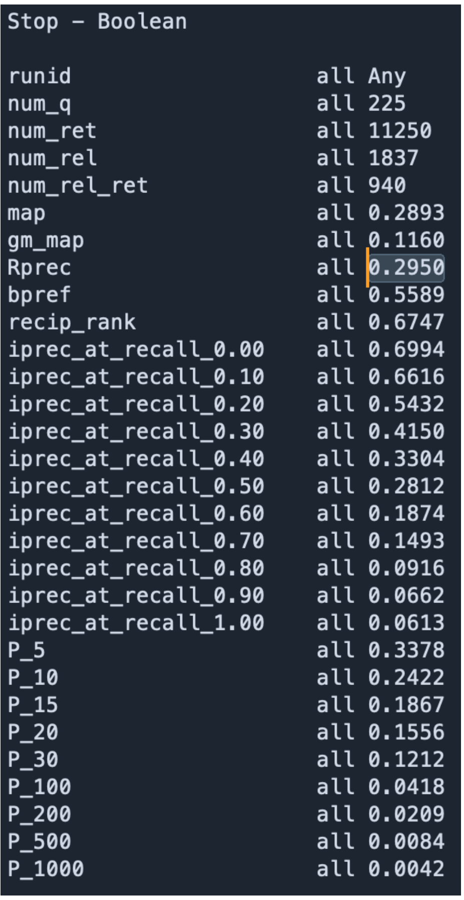
Stop Analyzer Scores



```
Stop Analyzer - getDefaultStopSet
Classic Similarity

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1075
map                     all 0.3645
gm_map                  all 0.2054
Rprec                   all 0.3585
bpref                   all 0.6363
recip_rank              all 0.7795
iprec_at_recall_0.00    all 0.7959
iprec_at_recall_0.10    all 0.7545
iprec_at_recall_0.20    all 0.6419
iprec_at_recall_0.30    all 0.5331
iprec_at_recall_0.40    all 0.4363
iprec_at_recall_0.50    all 0.3745
iprec_at_recall_0.60    all 0.2660
iprec_at_recall_0.70    all 0.2077
iprec_at_recall_0.80    all 0.1304
iprec_at_recall_0.90    all 0.0906
iprec_at_recall_1.00    all 0.0834
P_5                     all 0.3973
P_10                    all 0.2773
P_15                    all 0.2139
P_20                    all 0.1809
P_30                    all 0.1385
P_100                   all 0.0478
P_200                   all 0.0239
P_500                   all 0.0096
P_1000                  all 0.0048
```

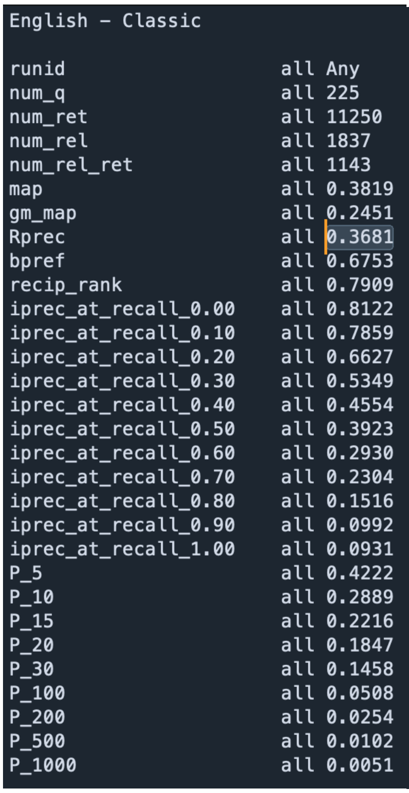


```
Stop - BM25

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1089
map                     all 0.3918
gm_map                  all 0.2136
Rprec                   all 0.3873
bpref                   all 0.6450
recip_rank              all 0.7964
iprec_at_recall_0.00    all 0.8090
iprec_at_recall_0.10    all 0.7722
iprec_at_recall_0.20    all 0.6642
iprec_at_recall_0.30    all 0.5597
iprec_at_recall_0.40    all 0.4709
iprec_at_recall_0.50    all 0.4160
iprec_at_recall_0.60    all 0.3027
iprec_at_recall_0.70    all 0.2303
iprec_at_recall_0.80    all 0.1521
iprec_at_recall_0.90    all 0.1066
iprec_at_recall_1.00    all 0.0982
P_5                     all 0.4347
P_10                    all 0.2982
P_15                    all 0.2287
P_20                    all 0.1898
P_30                    all 0.1430
P_100                   all 0.0484
P_200                   all 0.0242
P_500                   all 0.0097
P_1000                  all 0.0048
```

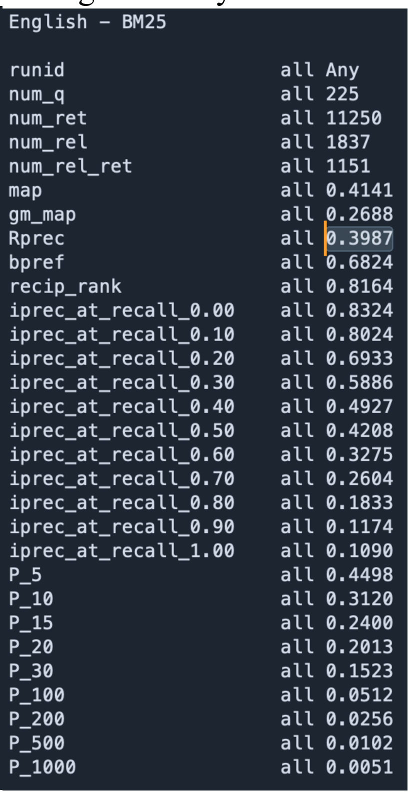


```
Stop - Boolean

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 940
map                     all 0.2893
gm_map                  all 0.1160
Rprec                   all 0.2950
bpref                   all 0.5589
recip_rank              all 0.6747
iprec_at_recall_0.00    all 0.6994
iprec_at_recall_0.10    all 0.6616
iprec_at_recall_0.20    all 0.5432
iprec_at_recall_0.30    all 0.4150
iprec_at_recall_0.40    all 0.3304
iprec_at_recall_0.50    all 0.2812
iprec_at_recall_0.60    all 0.1874
iprec_at_recall_0.70    all 0.1493
iprec_at_recall_0.80    all 0.0916
iprec_at_recall_0.90    all 0.0662
iprec_at_recall_1.00    all 0.0613
P_5                     all 0.3378
P_10                    all 0.2422
P_15                    all 0.1867
P_20                    all 0.1556
P_30                    all 0.1212
P_100                   all 0.0418
P_200                   all 0.0209
P_500                   all 0.0084
P_1000                  all 0.0042
```

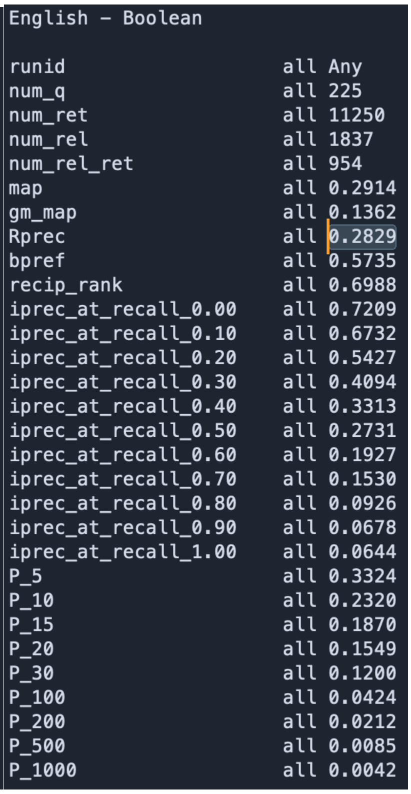
English Analyzer Scores



```
English - Classic

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1143
map                     all 0.3819
gm_map                  all 0.2451
Rprec                   all 0.3681
bpref                   all 0.6753
recip_rank              all 0.7909
iprec_at_recall_0.00    all 0.8122
iprec_at_recall_0.10    all 0.7859
iprec_at_recall_0.20    all 0.6627
iprec_at_recall_0.30    all 0.5349
iprec_at_recall_0.40    all 0.4554
iprec_at_recall_0.50    all 0.3923
iprec_at_recall_0.60    all 0.2930
iprec_at_recall_0.70    all 0.2304
iprec_at_recall_0.80    all 0.1516
iprec_at_recall_0.90    all 0.0992
iprec_at_recall_1.00    all 0.0931
P_5                     all 0.4222
P_10                    all 0.2889
P_15                    all 0.2216
P_20                    all 0.1847
P_30                    all 0.1458
P_100                   all 0.0508
P_200                   all 0.0254
P_500                   all 0.0102
P_1000                  all 0.0051
```



```
English - BM25

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 1151
map                     all 0.4141
gm_map                  all 0.2688
Rprec                   all 0.3987
bpref                   all 0.6824
recip_rank              all 0.8164
iprec_at_recall_0.00    all 0.8324
iprec_at_recall_0.10    all 0.8024
iprec_at_recall_0.20    all 0.6933
iprec_at_recall_0.30    all 0.5886
iprec_at_recall_0.40    all 0.4927
iprec_at_recall_0.50    all 0.4208
iprec_at_recall_0.60    all 0.3275
iprec_at_recall_0.70    all 0.2604
iprec_at_recall_0.80    all 0.1833
iprec_at_recall_0.90    all 0.1174
iprec_at_recall_1.00    all 0.1090
P_5                     all 0.4498
P_10                    all 0.3120
P_15                    all 0.2400
P_20                    all 0.2013
P_30                    all 0.1523
P_100                   all 0.0512
P_200                   all 0.0256
P_500                   all 0.0102
P_1000                  all 0.0051
```



```
English - Boolean

runid                   all Any
num_q                   all 225
num_ret                 all 11250
num_rel                 all 1837
num_rel_ret             all 954
map                     all 0.2914
gm_map                  all 0.1362
Rprec                   all 0.2829
bpref                   all 0.5735
recip_rank              all 0.6988
iprec_at_recall_0.00    all 0.7209
iprec_at_recall_0.10    all 0.6732
iprec_at_recall_0.20    all 0.5427
iprec_at_recall_0.30    all 0.4094
iprec_at_recall_0.40    all 0.3313
iprec_at_recall_0.50    all 0.2731
iprec_at_recall_0.60    all 0.1927
iprec_at_recall_0.70    all 0.1530
iprec_at_recall_0.80    all 0.0926
iprec_at_recall_0.90    all 0.0678
iprec_at_recall_1.00    all 0.0644
P_5                     all 0.3324
P_10                    all 0.2320
P_15                    all 0.1870
P_20                    all 0.1549
P_30                    all 0.1200
P_100                   all 0.0424
P_200                   all 0.0212
P_500                   all 0.0085
P_1000                  all 0.0042
```