Index Elimination

I tested the RankedQueryParser with no comparison and a comparison of $w_{q,t} < 0$. There was no significant computation time difference, so I decided to include index elimination threshold checking in the original RankedQueryParser class as there would be no noticeable change in search engine response time for default queries. This also makes the code simpler.

After testing the Cranfield corpus with different thresholds, the thresholds with the highest mean average precision were around 0.9. However, the improvements to MAP were small. The higher the threshold, the lower the mean response time (or higher throughput). However, the difference between the throughputs using each threshold was minimal when running only 30 queries. There might be a significant difference at a higher number of queries. Note, I was using the VS studio debugger, so the MRT differences might be even smaller when running the search engine through the command line.

Cranfield Test:

Default (No comparison):
MAP = 0.3476916638515977
MRT = 0.03366305881076389
throughput = 29.706153728378546

Default (Threshold = 0):
MAP = 0.3476916638515977
MRT = 0.033535503811306426
throughput = 29.819143485265073

Threshold = 1:
MAP = 0.3473509145611467
MRT = 0.03160350269741482
throughput = 31.642062260453187

Threshold = 2:
MAP = 0.2982170642523871
MRT = 0.030025500191582573
throughput = 33.305023850371775

Threshold = .95:
MAP = 0.34686139450985926
MRT = 0.031021497514512802
throughput = 32.23571007596166

Threshold = .94
MAP = 0.34686139450985926

MRT = 0.03125753084818522
throughput = 31.992290269404275

Threshold = .93
MAP = 0.3488252923754021
MRT = 0.03200987180074056
throughput = 31.240362542684867

Threshold = .92:
MAP = 0.3488252923754021
MRT = 0.03172041893005371
throughput = 31.525434837575354

Threshold = .91:
MAP = 0.3488252923754021
MRT = 0.031093933317396377
throughput = 32.16061441286111

Threshold = .9:
MAP = 0.3488252923754021
MRT = 0.03140732129414876
throughput = 31.839709940061066

Threshold = .89:
MAP = 0.3488252923754021
MRT = 0.03217898368835449
throughput = 31.07618343962485

Threshold = .87
MAP = 0.3476916638515977
MRT = 0.033719470765855575
throughput = 29.65645596705517

Threshold = .85:
MAP = 0.3480466085623333
MRT = 0.032003648546006944
throughput = 31.24643737299036

Threshold = .8:
MAP = 0.3480466085623333
MRT = 0.032310768763224286
throughput = 30.94943383514253

Threshold = .5:

MAP = 0.3476916638515977
MRT = 0.03395783848232693
throughput = 29.448281889921866

Threshold = 10:

MAP = 0.0
MRT = 0.029681552251180014
throughput = 33.69096034929387

Relevance Parks:

There are not enough queries and relevance sets for the Parks corpus to find a meaningful threshold.