1. Read Queries and Write Rankings

For bm25:

```
enwiki: Political %20status%20of%20Transnistria Q0 6019abc315e3afd5250d01a8897bee49b1646249 1 16.003521 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 ab1b4b7b8281f43fa23e3d16880c9e36483291cb 2 14.552347 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 2b66112b9092a962ccd9b32870396df2c2d3a501 3 14.236789 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 369765ebd82deed7c6ba3692d5c7d474a4193fbf 4 10.9345665 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 94715da76a2657b528bcef62af1cf5b4fe163815 5 10.194582 team2−standard
```

For custom scoring function:

```
enwiki: Political %20status%20of%20Transnistria Q0 2b266a28281a4c8cb0de3770e534182d2a15e999 1 5.0 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 79c1315ad446856d05cff6325f2f3dd52ec60900 2 5.0 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 56df8c80211225e9794d0b2ac6232df4337ba526 3 4.0 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 6019abc315e3afd5250d01a8897bee49b1646249 4 4.0 team2−standard
enwiki: Political %20status%20of%20Transnistria Q0 82c159e70f2c4a14b8d583bdfd8bb3ae3494d788 5 4.0 team2−standard
```

2. Evaluation with trec_eval

Method	**MAP**	**RPREC**	**NDCG@20**
BM25	0.6016	0.5966	0.7696
Custom	0.5096	0.5259	0.6740

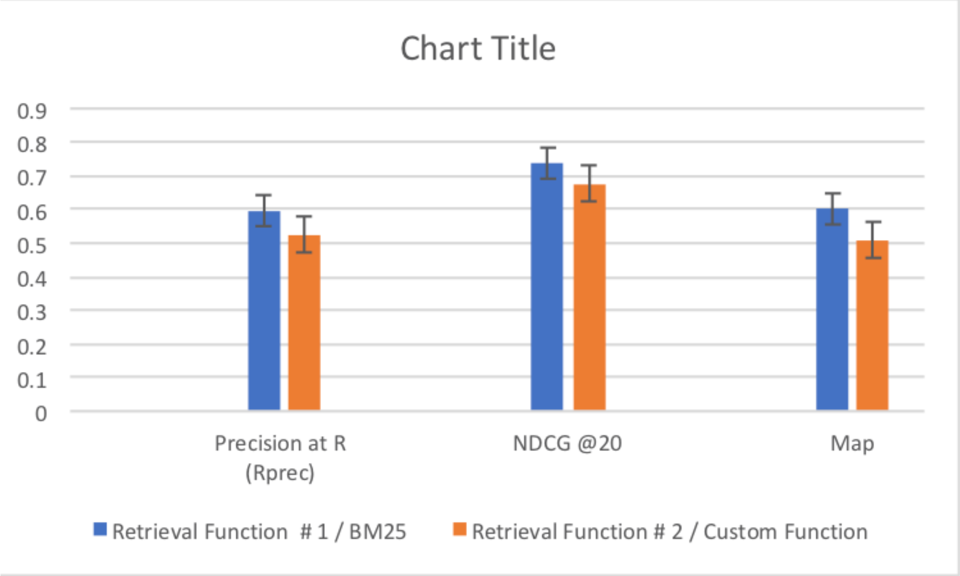
3. Our Implementations

Using our versions of MAP, RPREC, and NDCG@20, we obtained the following:

Method	**MAP**	**RPREC**	**NDCG@20**
BM25	0.60155	0.59637	0.76973
Custom	0.51017	0.52562	0.67322

The values seem very close. We are not sure what the difference could be (perhaps rounding error?). It could also be how we treat results where no documents were returned from a query: we treat them as 0's when we average.

4. Analysis of our Findings



Erm, please excuse the lack of creativity in the title... This is comparing the BM25 algorithm (in blue) with our custom scoring function (in orange).

While just looking at the bars we can see that BM25 is always higher in all metrics than the custom scoring function, we should also look at the error bars. In all cases, the error bars overlap. This typically means that the difference is not statistically significant. However, there is barely any overlap for MAP and the custom scoring function is really stupid, so let's just say that BM25 is better...