Comparison between the top 5 results from both the search engines:

(**Note**: The corpus has text files named from 1 to 1000 to avoid duplicate names. For the corresponding name of the document for the displayed document ID please refer KeyValueMapping.txt)

Query 1: hurricane isabel damage

Lucene:

1 Q0 3(doc_id = 223) 1 8.11566 Lucene

1 Q0 658(doc_id = 621) 2 8.002673 Lucene

1 Q0 656(doc_id = 619) 3 8.000477 Lucene

1 Q0 652(doc_id = 615) 4 7.956941 Lucene

1 Q0 664(doc_id = 628) 5 7.956941 Lucene

BM25:

1 Q0 corpus\780 1 1.8189060057681496 BM25

1 Q0 corpus\125 2 1.637461201773431 BM25

1 Q0 corpus\730 3 1.554134532280492 BM25

1 Q0 corpus\656 4 1.4374397741893858 BM25

1 Q0 corpus\742 5 1.4103580572615875 BM25

Among the top 5 results from both the search engines for this query we can see that just one document is common i.e

- 'Effects_of_Hurricane_Isabel_in_Pennsylvania' which is ranked 3rd in Lucene and 4th in BM25

This shows the rankings are skewed. The degree of overlap between the results of both the search engines is moderate.

##

Query 2: forecast models

Lucene:

2 Q0 278(doc_id = 199) 1 10.373764 Lucene

2 Q0 120(doc_id = 25) 2 9.826347 Lucene

2 Q0 277(doc_id = 198) 3 9.772875 Lucene

2 Q0 408(doc_id = 344) 4 9.73633 Lucene

2 Q0 83(doc_id = 812) 5 9.642991 Lucene

BM25:

2 Q0 corpus\408 1 4.080036295481349 BM25

2 Q0 corpus\278 2 3.723438278232643 BM25

2 Q0 corpus\120 3 3.695764422616924 BM25

2 Q0 corpus\83 4 3.561239754035285 BM25

2 Q0 corpus\277 5 3.554786432154346 BM25

All the top 5 ranked documents are common for both the search engines but in a different order for this query. Hence we can conclude that the overlap between the results of both the implementation is high.

- 'Tropical_cyclone_prediction_model' which is ranked 1st in Lucene and 2nd in BM25
- 'Tropical_cyclone_rainfall_forecasting' which is ranked 2nd in Lucene and 3rd in BM25
- 'Tropical_cyclone_track_forecasting' which is ranked 3rd in Lucene and 5th in BM25
- 'History_of_Atlantic_tropical_cyclone_warnings' which is ranked 4th in Lucene and 1st in BM25
- 'Weather_forecasting' which is ranked 5th in Lucene and 3rd in BM25

##

Query 3: green energy Canada

Lucene:

3 Q0 166(doc_id = 75) 1 7.2819915 Lucene

3 Q0 778(doc_id = 754) 2 7.1423182 Lucene

3 Q0 458(doc_id = 399) 3 7.0445404 Lucene

3 Q0 695(doc_id = 662) 4 6.9848013 Lucene

3 Q0 692(doc_id = 659) 5 6.2121935 Lucene

BM25:

3 Q0 corpus\458 1 2.466948511592372 BM25

3 Q0 corpus\177 2 2.4614616442187005 BM25

3 Q0 corpus\778 3 2.4130204100561325 BM25

3 Q0 corpus\692 4 2.4100154386578803 BM25

3 Q0 corpus\159 5 2.3063655419748574 BM25

Among the top 5 results from both the search engines for this query we can see that three documents is common but in different order i.e.

- 'Hydro-Qu%C3%A9bec' which is ranked 2nd in Lucene and 3rd in BM25
- 'Lake_Erie' which is ranked 5th in Lucene and 4th in BM25
- 'United_States_Department_of_Energy' which is ranked 3rd in Lucene and 1st in BM25

This shows the rankings are skewed. But this query is better than query 1. The degree of overlap between the results of both the search engines is high.

##

Query 4: heavy rains

Lucene:

4 Q0 802(doc_id = 782) 1 7.729933 Lucene

4 Q0 800(doc_id = 780) 2 7.6035967 Lucene

4 Q0 820(doc_id = 802) 3 7.4527225 Lucene

4 Q0 836(doc_id = 819) 4 7.340967 Lucene

4 Q0 829(doc_id = 811) 5 7.1677485 Lucene

BM25:

4 Q0 corpus\802 1 2.6471324627274564 BM25

4 Q0 corpus\800 2 2.622577095764745 BM25

4 Q0 corpus\508 3 2.615213200064224 BM25

4 Q0 corpus\820 4 2.5903657044701243 BM25

4 Q0 corpus\836 5 2.510142555888307 BM25

Among the top 5 results from both the search engines for this query we can see that four documents is common but in different order i.e.

- 'Hurricane_Hilda' which is ranked 1st in both Lucene and BM25
- 'Hurricane_Cleo' which is ranked 2nd in both Lucene and BM25
- 'Hurricane_Joan%E2%80%93Miriam' which is ranked 3rd in Lucene and 4th in BM25
- 'Hurricane_Michelle' which is ranked 4th in Lucene and 5th in BM25

The top 5 results from both the search engines are very similar. This query has a high overlap degree but is better than query 1 and 3.

##

Query 5: hurricane music lyrics

Lucene:

5 Q0 562(doc_id = 515) 1 10.68092 Lucene

5 Q0 580(doc_id = 535) 2 10.49033 Lucene

5 Q0 570(doc_id = 524) 3 9.936356 Lucene

5 Q0 584(doc_id = 539) 4 9.925853 Lucene

5 Q0 565(doc_id = 518) 5 9.696245 Lucene

BM25:

5 Q0 corpus\563 1 3.572249288048255 BM25

5 Q0 corpus\562 2 3.496240092635622 BM25

5 Q0 corpus\580 3 3.455896870638539 BM25

5 Q0 corpus\565 4 3.3693756441927847 BM25

5 Q0 corpus\584 5 3.26978940760694 BM25

Among the top 5 results from both the search engines for this query we can see that four documents is common but in different order i.e.

- 'Hurricane_(Natalie_Grant_album)' which is ranked 1st in Lucene and 2nd in BM25
- 'Badman_(EP)' which is ranked 2nd in Lucene and 3rd in BM25
- 'Addicted_Romantic' which is ranked 4th in Lucene and 5th in BM25
- 'Hurricane_(Athlete_song)' which is ranked 5th in Lucene and 4th in BM25

The top 5 results from both the search engines are very similar. This query has a high overlap degree but is better than query 1 and 3.

##

Query 6: accumulated snow

Lucene:

6 Q0 60(doc_id = 557) 1 8.306384 Lucene

6 Q0 65(doc_id = 612) 2 7.695259 Lucene

6 Q0 82(doc_id = 801) 3 7.4145412 Lucene

6 Q0 52(doc_id = 468) 4 7.148652 Lucene

6 Q0 49(doc_id = 434) 5 6.7129683 Lucene

BM25:

6 Q0 corpus\60 1 3.386132791443842 BM25

6 Q0 corpus\65 2 3.2200710788656193 BM25

6 Q0 corpus\82 3 3.0566409349094057 BM25

6 Q0 corpus\52 4 2.876400120159479 BM25

6 Q0 corpus\19 5 2.6967182484162606 BM25

Among the top 5 results from both the search engines for this query we can see that first four documents is common in the same order i.e.

- 'Graupel' which is ranked 1st in both Lucene and BM25
- 'Freezing_rain' which is ranked 2nd in both Lucene and BM25
- 'Severe_weather' which is ranked 3rd in both Lucene and BM25
- 'Winter_storm' which is ranked 4th in both Lucene and BM25

The top 5 results from both the search engines are almost similar. The overlap degree for this query is high.

##

Query 7: snow accumulation

Lucene:

7 Q0 68(doc_id = 645) 1 10.720123 Lucene

7 Q0 62(doc_id = 579) 2 9.909413 Lucene

7 Q0 52(doc_id = 468) 3 9.573895 Lucene

7 Q0 65(doc_id = 612) 4 9.360981 Lucene

7 Q0 96(doc_id = 956) 5 8.988663 Lucene

BM25:

7 Q0 corpus\68 1 4.497559463674471 BM25

7 Q0 corpus\62 2 4.185984116438346 BM25

7 Q0 corpus\52 3 4.022373947916401 BM25

7 Q0 corpus\65 4 3.9429873815840155 BM25

7 Q0 corpus\70 5 3.753226466027644 BM25

Among the top 5 results from both the search engines for this query we can see that first four documents is common in the same order i.e.

- 'Rain_and_snow_mixed' which is ranked 1st in both Lucene and BM25
- 'Ice_pellets' which is ranked 2nd in both Lucene and BM25
- 'Winter_storm' which is ranked 3rd in both Lucene and BM25
- 'Freezing_rain' which is ranked 4th in both Lucene and BM25

The top 5 results from both the search engines are almost similar. The overlap degree for this query is high.

##

Query 8: massive blizzards blizzard

Lucene:

8 Q0 55(doc_id = 501) 1 18.241514 Lucene

8 Q0 54(doc_id = 490) 2 17.44734 Lucene

8 Q0 52(doc_id = 468) 3 12.564058 Lucene

8 Q0 82(doc_id = 801) 4 11.573803 Lucene

8 Q0 37(doc_id = 301) 5 10.966423 Lucene

BM25:

8 Q0 corpus\55 1 7.812895495281669 BM25

8 Q0 corpus\54 2 7.396376644295811 BM25

8 Q0 corpus\52 3 5.40906372969876 BM25

8 Q0 corpus\82 4 4.9394278301377375 BM25

8 Q0 corpus\37 5 4.74244599729721 BM25

All the top 5 ranked documents for this query from both the search engines are the same. So we can say that the degree of overlap between these 2 set of results for this particular query is the highest.

Query 9: new york city subway

Lucene:

9 Q0 754(doc_id = 728) 1 11.8447485 Lucene

9 Q0 708(doc_id = 677) 2 11.180321 Lucene

9 Q0 956(doc_id = 952) 3 11.147384 Lucene

9 Q0 691(doc_id = 658) 4 10.743515 Lucene

9 Q0 472(doc_id = 415) 5 10.133621 Lucene

BM25:

9 Q0 corpus\708 1 4.064707699516038 BM25

9 Q0 corpus\956 2 3.8565293312143423 BM25

9 Q0 corpus\754 3 3.823492056316846 BM25

9 Q0 corpus\573 4 3.782754755425798 BM25

9 Q0 corpus\691 5 3.409399369650968 BM25

Four of the top 5 ranked documents are common for both the search engines but in a different order for this query.

- 'Baltimore,_Maryland' which is ranked 1st in Lucene and 3rd in BM25
- 'Washington_Metro' which is ranked 2nd in Lucene and 1st in BM25
- 'ISS_(disambiguation)' which is ranked 3rd in Lucene and 2nd in BM25
- 'Pittsburgh' which is ranked 4th in Lucene and 5th in BM25

Hence we can conclude that the overlap between the results of both the implementation is high.

##########

Conclusion :

The reason there is difference in the ranked result list between Lucene and BM25 is because of the underlying models implemented.

Lucene implements a combination of two models, Boolean model and Vector Space model. Lucene considers two factors before generating score for a particular document, query term frequency in a document and number of hits.
Boolean model handles the factor of number of hits. Number of hit gives the details of how many documents present within the corpus consists of the given query term. So based on the number of hits, Boolean model will assign the final rank to that particular document.

Vector space model handles the query term frequency factor. If a query term appears more in one document than other documents in the corpus then that document will be considered to be more relevant to the given query.

Whereas it is different in case of BM25, if a document has more number of hits then its ranking goes down. This is because BM25 uses probabilistic model to determine whether the document is relevant to the given query or not.

Because of these differences there are variations between the results of these two search engines.