Documents used for analysis

- https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html
- https://lucene.apache.org/core/3_5_0/scoring.html

Lucene's practical scoring formula is as follows:

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum (tf(t \text{ in } d) \cdot idf(t)2 \cdot t.getBoost() \cdot norm(t,d))$$

Where q -> query
d -> document
t -> each query term

Queries
1: global warming potential
2: green power renewable energy
3: solar energy california
4: light bulb bulbs alternative alternatives

Query 1: global warming potential

1 Q0 Globalwarming.txt 1 0.31433234 Lucene
1 Q0 UnitedNationsFrameworkConventiononClimateChange.txt 2 0.19791259 Lucene
1 Q0 Environmentalimpactoftheenergyindustry.txt 3 0.1885559 Lucene
1 Q0 Sustainabilityandenvironmentalmanagement.txt 4 0.172979 Lucene
1 Q0 Climatechangemitigation.txt 5 0.16933782 Lucene

1 Q0 Globalwarming.txt 1 5.52757435298 BM25
1 Q0 Climatechangemitigation.txt 2 5.40858485799 BM25
1 Q0 Environmentalimpactoftheenergyindustry.txt 3 5.40441247036 BM25
1 Q0 UnitedNationsFrameworkConventiononClimateChange.txt 4 5.35549067316 BM25
1 Q0 Naturalenvironment.txt 5 5.24985999618 BM25

There are 4 common documents in both the quries. Although the ranking differs we can see that the document ranking do not vary to a large extend.

Query 2: green power renewable energy

2 Q0 3Degrees.txt 1 0.4200997 Lucene
2 Q0 RenewableEnergyCertificate.txt 2 0.3626191 Lucene
2 Q0 RenewableEnergyCertificates.txt 3 0.3626191 Lucene
2 Q0 Greenjob.txt 4 0.34853685 Lucene
2 Q0 RenewableenergyinMexico.txt 5 0.33668244 Lucene

2 Q0 Sustainabilityandenvironmentalmanagement.txt 1 2.49855514244 BM25
2 Q0 Greenpaper.txt 2 2.44264179217 BM25
2 Q0 GreenBuildingMIT.txt 3 2.42229768148 BM25

2 Q0 Greenaccounting.txt 4 2.40765176323 BM25
2 Q0 Urbanhorticulture.txt 5 2.18321607142 BM25

These list of documents are for the query "green power renewable energy".
There are no common documents in the 2 sets of documents.
This happens because of the way the weights are normalized.
Normalization plays a huge role in this implementation because of the enoromous corpus size that is considered.

Query 3: solar energy california

3 Q0 NevadaSolarOne.txt 1 0.3151702 Lucene
3 Q0 SolarDecathlon.txt 2 0.29580134 Lucene
3 Q0 SiliconValleyPower.txt 3 0.2952835 Lucene
3 Q0 RenewableenergyintheUnitedStates.txt 4 0.29401615 Lucene
3 Q0 RenewableenergyinArmenia.txt 5 0.29293057 Lucene

3 Q0 KernCountyCalifornia.txt 1 3.62942294402 BM25
3 Q0 LosAngeles.txt 2 3.51844370919 BM25
3 Q0 Emissionstandard.txt 3 3.49478566539 BM25
3 Q0 CamarilloCalifornia.txt 4 3.46665954816 BM25
3 Q0 Exhaustgas.txt 5 3.23022806682 BM25

The scoring seen above are for the query "solar energy california".There exists no common documents among the top 5 results obtained for the mentioned query in both the retrieval models.

Query 4: light bulb bulbs alternative alternatives

4 Q0 Phaseoutofincandescentlightbulbs.txt 1 0.6076199 Lucene
4 Q0 Incandescentlightbulb.txt 2 0.3613186 Lucene
4 Q0 Incandescentlightbulbs.txt 3 0.3613186 Lucene
4 Q0 Energyconservation.txt 4 0.21741651 Lucene
4 Q0 Energysavinglamp.txt 5 0.1982192 Lucene

4 Q0 Phaseoutofincandescentlightbulbs.txt 1 27.4595978626 BM25
4 Q0 Incandescentlightbulb.txt 2 24.4033256696 BM25
4 Q0 Incandescentlightbulbs.txt 3 24.4033256696 BM25
4 Q0 Energyconservation.txt 4 21.0878209459 BM25
4 Q0 Compactfluorescentlamp.txt 5 20.8119487459 BM25

The scoring seen above are for the query "light bulb bulbs alternative alternatives"

There are 4 common documents among the top 5 results obtained for the mentioned query in both the retrieval models.

The change in scoring varies because of various factors that are computed to score.

Analysis

- The above tables shows the top 5 results obtained by both the retrieval system for the queries.
- We can see a significant change in the results of both the systems. This is Cleary because of the scoring implementation.
- The major reason for this boosting. Lucene implements the concept of boosting. Lucene might boost certain documents based on the source of those documents.

But my implementation does not have sort of boosting.

There is a significant difference in the way both the models calculate the values for each of the pages. It depends on the term frequency also. While lucene lucene multiplies the tf with idf thus causing considerable change in the weight of the document. If we have short documents or very long ones, BM25 should be better.