# Comparison of Results:

Lucene combines Boolean model (BM) of Information Retrieval with Vector Space Model (VSM) of Information Retrieval - documents "approved" by BM are scored by VSM.

BM25 extends the scoring function for the binary independence model to include document and query term weights. BM25 uses probabilistic model to determine whether the given query is relevant or not.

The document below is the top 5 ranked documents generated from Lucene and BM25;



**Query Q1:**

The ranked set of documents for query “hurricane isabel damage” is the same for both Lucene and BM25. The score for documents are same for ‘654’ and ‘666’ in both Lucene and BM25 and since the Lucene takes the docID into consideration, the document ‘654’ is ranked above ‘666’.

**Query Q2:**

The ranked set of documents for query “forecast models” is same for both models. The scores of the documents differ since both Lucene and BM25 have different scoring models.

**Query Q3:**

Among the top ranked documents for query “green energy Canada” the Lucene does not consist of document ‘275’ while BM25 does not consist of document ‘697’. The ranking is almost the same apart from this difference. The document ‘697’ is ranked higher in Lucene having the term frequency alone in account. i.e the document contains 17 occurrences of term ‘canada’ of query.

**Query Q4:**

The ranked set of documents for query “heavy rains” is same for both models. The scores of the documents differ since both Lucene and BM25 have different scoring models.

**Query Q5:**

The ranked set of documents for query “hurricane music lyrics” is almost the same for both the models with the difference in just the 5th rank. But if you further go down the list of ranked documents the 6th ranked document according to BM25 is ‘552’, hence it implies the ranking is similar in both cases.

**Query Q6:**

The ranked set of documents for query “accumulated snow” is almost the same for both the models with the difference in just the 5th rank. But if you further go down the list of ranked documents the 6th ranked document according to BM25 is ‘49’, hence it implies the ranking is similar in both cases.

**Query Q7:**

The ranked set of documents for query “snow accumulation” is almost the same for both the models with the difference in just the 5th rank. But if you further go down the list of ranked documents the 6th ranked document according to BM25 is ‘96’, hence it implies the ranking is similar in both cases.

Also the query Q7 has some overlap of the top ranked documents with query Q6 because of the similarity in root words and presence of a common query token term “snow”.

**Query Q8:**

The ranked set of documents for query “massive blizzards blizzard” is almost the same for both the models with the difference in just the 5th rank. But if you further go down the list of ranked documents the 6th ranked document according to BM25 is ‘37’ and in Lucene the 6th ranked document is ‘78’, hence it implies the ranking is similar in both cases.

**Query Q9:**

The ranked documents for query “new york city subway” are almost similar for the top 5 ranked documents but sorted in different order. The overlap is high between both the models.

**CONCLUSION**:

Since the relevance information is not taken into account in BM25 the ranking of documents based on the given queries is almost similar for both Lucene and BM25. Once relevance information is considered we may see the variation in the documents ranking and scores.