For Query1 “global warming potential” gives the following documents as Top5:

|  |  |
| --- | --- |
| Task1(Lucene) | Task2(VSM) |
| Global warming | Global warming |
| Sustainability and environmental management | Joseph Romm |
| Climate change mitigation | Climate change mitigation |
| Environmental impact of the energy industry | Environmental impact of the energy industry |
| Climate change | Sustainability and environmental management |

The top 5 documents from Task2 using cosine similarity include the documents with terms “global”, “warming” and “potential”. The page “Global Warming” contains all the 3 terms very frequently and is also a fairly large document, hence it appears at the top of the list for both retrieval modules. The VSM does not consider the Boolean model logic, which is why the page “Jospeh Romm” appears in the vsm as it is also a small document, so its weightage is more in our case. Lucene uses Boolean model as well, hence the ranking of this page is lower for Lucene. The rest of the pages in the list have all the query terms fairly frequently and are also small documents and thus appear in both modules as top5.

No stemming and stopping was done for the corpus in Task2. Also a lot of processing was done on the text to create index, hence the top documents would be those taking into consideration the length of documents.

For Query2 “green power renewable energy” gives the following documents as Top5:

|  |  |
| --- | --- |
| Task1 (Lucene) | Task2 (VSM) |
| 3Degrees | Renewable energy in Mexico |
| Green job | Renewable energy payments |
| Renewable energy in Mexico | 10025renewableenergy |
| World Energy Engineering Congress | Renewable energy in Luxembourg |
| Renewable Energy Certificate | Renewable energy in NewZealand |

The top page, “3Dregrees” contains the query terms frequently and it is ranked as the most similar document for the query by Lucene. The page “Renewable energy in Mexico” contains the terms “renewable” and “energy” very frequently taking into consideration the size of the document. Hence it is ranked as top most in VSM. However it does not contain all the query terms that frequently, which is why it is ranked a little lower by Lucene because it considers pages containing all the query terms to have more weight.

The other documents from VSM have most of the query terms except one or two quite frequently, hence it is scored better by VSM as compared to Lucene. The top documents from Lucene are the ones with all query terms at least once and ranked as per the frequency of query terms.

For Query3 “solar energy california” gives the following documents as Top5:

|  |  |
| --- | --- |
| Task1 (Lucene) | Task2 (VSM) |
| NevadaSolarOne | Solar energy |
| RenewableenergyintheUnitedStates | Nevada Solar One |
| SolarDecathlon | California |
| Solarenergy | Renewable energy in Asia |
| Solarpower | Renewable energy in the United States |

The top document in VSM is the page which contains “solar” “energy” very frequently but “california less frequently. However the top document from Lucene has “California” more frequent than VSM.

The second top most document in VSM contains query terms except “California” frequently, hence it comes in the top 2. However Lucene gives more weight to documents containing all terms hence it occurs a little below in the top documents in Lucene.

The rest of the top documents contain the query terms frequently and are not very large.

For Query4 “light bulb bulbs alternative alternatives” gives the following documents as Top5:

|  |  |
| --- | --- |
| Task1 (Lucene) | Task2 (VSM) |
| Phase out of incandescent light bulbs | Phase out of incandescent light bulbs |
| Energy saving lamp | Incandescent light bulbs |
| Incandescent light bulb | Incandescent light bulb |
| Incandescent light bulbs | Energy saving lamp |
| Energy conservation | LED lights |

The top documents for this query are almost matching by VSM as well as Lucene because they contain all the query terms frequently and are small documents, hence their score is large.

Thus, the documents ranked by Lucene are similar to those ranked by VSM in some cases.

Lucene uses a combination of VSM and Boolean model, hence the scores calculated for documents differ as weightage for documents which strictly contain all the query terms is more. VSM does not implement this logic, hence the top documents may differ.

Lucene also applies a normalization formula for the documents and calculates the scores a little differently by using TopScoreDocCollector. In VSM, score of the document is calculated on the basis of the term frequency and the relative size of the document compared to all documents in the corpus. VSM gives a good output when we want to get documents which may contain any of the query terms frequently.

Thus, the top documents using a combination of Boolean model and VSM appear to be a better choice for practical purposes where the user wants all the query terms to be present in his results.