Referred document of lucene scoring https://lucene.apache.org/core/3\_6\_0/scoring.html

lucene scoring implements Boolean retrieval model and vector space model to determine how relevant the document is to a given query. It uses the Boolean model to first narrow down the documents that need to be scored based on the use of boolean logic in the Query specification.

It then implements the vector space implementation to determine the similarity. It also uses query boosting techniques and length normalization technique.

My implementation does not normalize the term frequency. I include the frequency of the term as it as in the weight calculation. The idf is log(n/df) where df is document frequency and n is the number of documents in the corpus. The magnitude of the document is the sqrt(sum of squares of the tf\*idfs of all the terms in the document). The idf for query term is considered as 1 as queries provided are short queries and normalizing them will not bring in any significant change.

Query 1 “global warming potential” comparison

Task 2 results

1,Q0,Globalwarming.txt,0.33627825557348356,COSINE\_SIM  
1,Q0,Climatechangemitigation.txt,0.14444884726645163,COSINE\_SIM  
1,Q0,Environmentalissue.txt,0.13585922313994797,COSINE\_SIM  
1,Q0,Environmentalimpactoftheenergyindustry.txt,0.12670578897925322,COSINE\_SIM  
1,Q0,UnitedNationsGlobalCompact.txt,0.11390543780082768,COSINE\_SIM

Task1 results

1,Q0,Globalwarming.txt,0.32637542,lucence\_indexing  
1,Q0,Environmentalimpactoftheenergyindustry.txt,0.18664117,lucence\_indexing  
1,Q0,Adaptationtoglobalwarming.txt,0.17725246,lucence\_indexing  
1,Q0,Sustainabilityandenvironmentalmanagement.txt,0.17119312,lucence\_indexing  
1,Q0,Climatechangemitigation.txt,0.16996518,lucence\_indexing

As we can notice the

Globalwarming.txt,

Climatechangemitigation.txt and

Environmentalimpactoftheenergyindustry.txt documents are the common documents listed in top 5. All these documents contain the 3 query terms. The discrepancies in the result could be because lucene uses query boosting and length normalization. Lucene fetches for the exact match for the query assigns higher precedence.

Query 2 “green power renewable energy” comparison

Task 2 results

2,Q0,Sustainableenergy.txt,0.4885296327204429,COSINE\_SIM  
2,Q0,Renewableenergycommercialization.txt,0.42595376040291755,COSINE\_SIM  
2,Q0,Renewableelectricity.txt,0.42453780741306785,COSINE\_SIM  
2,Q0,Renewableenergy.txt,0.42453780741306785,COSINE\_SIM  
2,Q0,100%25renewableenergy.txt,0.4014715473605057,COSINE\_SIM

Task1 results

2,Q0,RenewableenergyintheCzechRepublic.txt,0.5122485,lucence\_indexing  
2,Q0,3Degrees.txt,0.46456742,lucence\_indexing  
2,Q0,RenewableenergyinMexico.txt,0.45119154,lucence\_indexing  
2,Q0,RenewableenergyinIndia.txt,0.43550903,lucence\_indexing  
2,Q0,RenewableEnergyCertificate.txt,0.4220634,lucence\_indexing

Lucene scoring has ranked the documents which have exact match of the query whereas my implementation looks for occurrence of the unigrams from the query.

Query 3 “solar energy california” comparison

Task2 results

3,Q0,Solarenergy.txt,0.456924831912427,COSINE\_SIM  
3,Q0,RenewableenergyintheUnitedStates.txt,0.43226085543992243,COSINE\_SIM  
3,Q0,California.txt,0.3987576266869611,COSINE\_SIM  
3,Q0,RenewableenergyinAsia.txt,0.38491927894313316,COSINE\_SIM  
3,Q0,Renewableelectricity.txt,0.3605250591777366,COSINE\_SIM

Task1 results

3,Q0,NevadaSolarOne.txt,0.39287785,lucence\_indexing  
3,Q0,SolarDecathlon.txt,0.3650384,lucence\_indexing  
3,Q0,Solarenergy.txt,0.3515552,lucence\_indexing  
3,Q0,RenewableenergyintheUnitedStates.txt,0.35039952,lucence\_indexing  
3,Q0,Listofrenewableenergytopicsbycountry.txt,0.3408028,lucence\_indexing

As we can notice here,

Solarenergy,

RenewableenergyintheUnitedStates,

Are the common queries in the top 5, the order is different as my approach does not give importance to the occurrence of bigrams and trigrams whereas lucene scoring is doing that.

My implementation assumes the importance of each word equally and the relevance is considered individually.

Query 4 “light bulb bulbs alternative alternatives” comparison

Task2 results

4,Q0,Alternativeenergy.txt,0.1122628168317465,COSINE\_SIM  
4,Q0,Rationalplanningmodel.txt,0.10657766704814639,COSINE\_SIM  
4,Q0,Energyconservation.txt,0.09994802453275296,COSINE\_SIM  
4,Q0,Darkskymovement.txt,0.07440540834194506,COSINE\_SIM  
4,Q0,Kilowatthour.txt,0.07020540406890342,COSINE\_SIM

Task1 results

4,Q0,Energyconservation.txt,0.25279185,lucence\_indexing  
4,Q0,Greenmarketing.txt,0.17730784,lucence\_indexing  
4,Q0,RenewableenergyinPakistan.txt,0.09172754,lucence\_indexing  
4,Q0,Energydevelopment.txt,0.08940669,lucence\_indexing  
4,Q0,Kilowatthour.txt,0.08856681,lucence\_indexing

The common documents are

Energyconservation,

Kilowatthour

The scores for these files and the order are matching the weight of the terms light bulb bulbs is matching the with what lucence calculates.