Report of Cosine Similarity and Relevance

Net Id: kxk152430
Homework - 2

Lemmatization software/code/library you used:
stanford-corenlp-3.3.1-models.jar
stanford-corenlp-3.3.1-models.jar

URL to download :
https://repository.cloudera.com/artifactory/repo/edu/stanford/nlp/stanford-corenlp/3.3.1/

---

1. Turn in the vector representation of the query (10 points per weighting scheme), and the top 5 documents for the query under both weighting schemes (50 points, with 25 points per weighting scheme). You are also required to present the vector representations for each of the first 5 ranked documents.

Included in ReadMe Output

2. Indicate the rank, score, external document identifier, and headline, for each of the top 5 documents for each query.

Included in ReadMe Output

3. Identify which documents you think are relevant and non-relevant for each query.
4. Describe why the top-ranked non-relevant document for each query did not get a lower score.

Q1 :
what similarity laws must be obeyed when constructing aeroelastic models
of heated high speed aircraft

Doc ID	Relevance(Y/N)	Reason
875	Yes	Because of title of the doc has it
429	Yes	High speed is present
509	Yes	Temperature and Heat is present
795	No	Just speed is present
141	Yes	High speed in title

Q2 :
what are the structural and aeroelastic problems associated with flight
of high speed aircraft

Doc ID	Relevance(Y/N)	Reason
875	Yes	High speed in title
429	Yes	High speed is present

896	No	Just high is present
12	Yes	High speed is present
650	No	Just problem is present

Q3 :
what problems of heat conduction in composite slabs have been solved so far

Doc ID	Relevance(Y/N)	Reason
485	Yes	Heat is present in title
181	No	Problem is in content
281	No	Conduction is present in content
119	Yes	Heat conduction is present
1283	No	Composite is present

Q4 :
can a criterion be developed to show empirically the validity of flow solutions for chemically reacting gas mixtures based on the simplifying assumption of instantaneous local chemical equilibrium

Doc ID	Relevance(Y/N)	Reason
437	Yes	Chemical and equilibrium are present
855	No	Flow is just present
939	No	Mixtures are present in content
1061	Yes	Local chemical is present
1189	Yes	Criterion is present

Q5 :
what chemical kinetic system is applicable to hypersonic aerodynamic problems

Doc ID	Relevance(Y/N)	Reason
471	Yes	Hypersonic and aerodynamic is present in title
995	No	Chemical is present in content
567	No	Aerodynamic is alone present
458	Yes	Hypersonic and problems are present
540	No	Chemical alone is present

Q6 :what theoretical and experimental guides do we have as to turbulent
couette flow behaviour

Doc ID	Relevance(Y/N)	Reason
385	No	Turbulent is available in content
271	Yes	Flow behavior is present
137	Yes	Theoritical is present
339	No	behavior alone
291	Yes	Turbulent and couetter is present

Q7 :
is it possible to relate the available pressure distributions for an
ogive forebody at zero angle of attack to the lower surface pressures of
an equivalent ogive forebody at angle of attack

Doc ID	Relevance(Y/N)	Reason
492	Yes	Pressure and distributions are present
248	No	Surface alone
57	Yes	Surface and pressure are present
1006	Yes	Equivalent, distributions are present in title
56	No	Lower is present

Q8 :
what methods -dash exact or approximate -dash are presently available
for predicting body pressures at angle of attack

Doc ID	Relevance(Y/N)	Reason
248	No	Attack is present in content
556	Yes	Body pressures is present tin title
122	Yes	Approximate, body, angle is present
492	No	Only angle is present
69	No	Pressure is alone there

Q9 :
papers on internal /slip flow/ heat transfer studies

Doc ID	Relevance(Y/N)	Reason
21	Yes	Heat transfer studies is present
875	No	Flow is alone there
963	No	Papers flow is there

436	Yes	Heat tranfer is there
509	No	Studies alone present

Q10 :are real-gas transport properties for air available over a wide range of enthalpies and densities

Doc ID	Relevance(Y/N)	Reason
436	Yes	Densities, enthalpies are present
405	No	Real gas is alone there
609	Yes	Transport properties are there in content
437	No	Wide Range is there
362	Yes	Densities, real gas is there

Q11 :
is it possible to find an analytical, similar solution of the strong blast wave problem in the newtonian approximation

Doc ID	Relevance(Y/N)	Reason
495	No	Approximate is alone present
609	Yes	Analytic solution is there
939	Yes	Wave problem is present in title
258	No	Solution is only there
320	Yes	Approximate analytic solution is there

Q12 :
how can the aerodynamic performance of channel flow ground effect machines be calculated

Doc ID	Relevance(Y/N)	Reason
650	Yes	Ground effect machines is present
939	No	Aerodynamic is alone there
592	Yes	Channel and flow are there
1132	Yes	Calculated is there
137	No	Aerodynamic, channel, flow is there

Q13 :
what is the basic mechanism of the transonic aileron buzz

Doc ID	Relevance(Y/N)	Reason
496	Yes	transonic is in title
880	Yes	Mechanism aileron is present
258	No	Mechanism alone there
795	No	Transonic is alone there
38	No	Basic mechanism is alone there

Q14 :
papers on shock-sound wave interaction

Doc ID	Relevance(Y/N)	Reason
291	Yes	Shock sound is available in title
875	Yes	Papers interaction is there
609	Yes	Sound wave is there
939	No	Papers alone present
1276	No	Interaction alone there

Q15 :
material properties of photoelastic materials

Doc ID	Relevance(Y/N)	Reason
462	Yes	Photoelastic is in title
1043	No	Material is alone there
1099	Yes	Material properties both there
1025	No	Photoelastic is alone there
542	No	Material alone present

Q16 :
can the transverse potential flow about a body of revolution be
calculated efficiently by an electronic computer

Doc ID	Relevance(Y/N)	Reason
161	Yes	Electronic, computer both present
1358	Yes	Transverse revolution there
106	No	Flow alone there
609	No	Electronic alone present
920	Yes	Efficiently, calculated, computer are there

Q17 :
can the three-dimensional problem of a transverse potential flow about
a body of revolution be reduced to a two-dimensional problem

Doc ID	Relevance(Y/N)	Reason
106	Yes	Three-dimensional is in title
372	Yes	Transverse potential flow is present
1301	No	Revolution, body is present
916	No	Three-dimensional, potential is there in content
1108	Yes	Problem revolution flow are present

Q18 :
are experimental pressure distributions on bodies of revolution at angle
of attack available

Doc ID	Relevance(Y/N)	Reason
291	Yes	Experimental pressure distributions are present in content
161	Yes	Attack angle is present
609	No	Pressure distributions are there alone
248	No	Revolution alone there
1005	Yes	Experimental, bodies are present

Q19 :
does there exist a good basic treatment of the dynamics of re-entry
combining consideration of realistic effects with relative simplicity of
results

Doc ID	Relevance(Y/N)	Reason
471	Yes	Combining, realistic, results are there
995	No	Realistic is alone there
458	No	Consideration, relative are alone there
519	Yes	Dynamics, re-entry are there
959	Yes	Simplicity combining consideration are all there

Q20 :
has anyone formally determined the influence of joule heating, produced
by the induced current, in magnetohydrodynamic free convection flows
under general conditions

Doc ID	Relevance(Y/N)	Reason
495	Yes	Magnet dynamics are in title
881	Yes	Joule heat is all there
58	No	Free flows are alone present
875	No	Hydrodynamic is there
30	No	Induced current is alone there

5. Briefly discuss the different effects you notice with the two weighting schemes, either on a query-by-query basis or overall, whichever is most illuminating. For example, you can point out that the weighting scheme seems to be working for this query as well as a list of other queries, but not for some other queries you have noticed. Try to explain why it works and why it does not work.

The documents were not relevant exactly, but they had terms of the query and so it resulted in more weight even if the document is not completely relevant. But in real case, the weighting schemes should result in giving more weight to actual relevant documents based on the meaning of context but not literal meaning. Hence based on the results, Weighting scheme 1 gave better results than Okapi weighting.
Also the Okapi weighting had doc length and average doc length which is not the criteria for exact meaningful matching context. To get good results the weighting scheme should take main consideration of term frequencies and doc frequencies and not lengths. The terms that are high frequent secured less weights in the second one.

6. Describe the design decisions you made in building your ranking system.

Tokenize.java : Performs tokenizing task, which removes unwanted characters, case folding, removing numbers, etc. This also does removal of stop words.
The stop words list taken from
/people/cs/s/sanda/cs6322/resourcesIR

Lemmatize.java: Performs lemmatization using the above mentioned standford nlp library, the dictionary and the posting lists are created based on that. The data structure doc frequency, posting list pointer and term pointer and the long dictionary string is all present and build inside this by calling the compression techniques functions.

DictionaryClass.java : This is a POJO class that renders the data structure for dictionary and pointer to posting list.

DocDetails.java : This is a POJO class that renders the data structure for posting list.