Relevance and Ranked Retrieval

Diksha Sharma (dxs134530)
CS 6322.001 – Information Retrieval – fall 2015
Assignment 3 - Report

APPROACH FOLLOWED

In this assignment I have used TreeMaps and replaced my old token using ArrayList with a TreeMap.

Class description is as below:

1. Search – This contains the main method and in turn calls other classes for execution.
2. Tokenizer – This creates the token index – (Part of assignment 1) – Changed array lists to hash maps. Stop words are stored.
3. Lemmatize – This class finds lemma for tokens in the token index and creates an index of lemmas. No Stop Words are stored. We find the lemma using the Standford Core NLP lemmatizer. Once the index is created – this class makes call for its compression in the Compress class. This class also creates the stem index and then calls compression method in Compress class for stem index compression. This class now is modified to do the ranked retrieval for the index created for the lemmas in document. This class also performs the weighting schemes on queries and documents and then peformed ranked retrieval on both weighting schemes.
4. Compress – This class does all the compression of both the indexes for tokens and stems. I use treemap for storing the indexes. The uncompressed indexes use Term as structure for storing the posting files. The posting file structure is a TreeMap. The compressed indexes use CompressToken and CompressStem for storing the posting files. The posting files are stored in array lists for terms of block 8.
5. PostingFile – This stores the document id and the frequency of the term/stem in that document.
6. CompressStem – Data Structure for compressed stem index used as a TreeMap.
7. CompressToken – Data Structure for compressed token index used as a treemap.
8. Porter – Porter stemmer – modified to create index with document id and frequency information for each term.
9. Document – Data Structure used for storing document information.
10. Term - Data structure used for uncompressed indexes.
11. Lemma – Data Structure used for storing lemma information for documents.
12. Query - Data Structure used for storing query information.

ANSWERS TO QUESTIONS IN ASSIGNMENT

FOR each query:

- **Turn in the vector representation of the query.**

  In folder: Query Vectors

- **The top 5 documents for the query under both weighting schemes. You are also required to present the vector representations for each of the first 5 ranked documents.**

  In folder: Ranked Retrieval/Max Tf Term Weight and Ranked Retrieval/Okapi Weight

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

  In folder: Ranked Retrieval/Max Tf Term Weight and Ranked Retrieval/Okapi Weight

- **Identify which documents you think are relevant and non-relevant for each query. Describe why the top-ranked non-relevant document for each query did not get a lower score. Briefly discuss the different affects 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.**

- **Query 1: what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft**

Max TF Term Weight:

Document Id	**Relevant (Yes/No)**	
51	No	Ranked highest but looks non relevant based on its text. The reason it is ranked higher for Q1 is because of high weight for term “aircraft” in both document and query.
573	No	Irrelevant but ranked high due to presence of “similitude”
486	Yes	
184	Yes	Irrelevant but ranked high due to presence of “similarity”, “aeroelastic” and other common words between the query and document.
13	Yes	

Okapi Weight:

Document Id	**Relevant (Yes/No)**	
51	No	Ranked highest but looks non relevant based on its text. The reason it is ranked higher for Q1 is because of high weight for term “aircraft” in both document and query.
486	Yes	

573	No	Ranked high due to presence of lemmas like “similitude”
878	No	
12	Yes	

- **Query 2: what are the structural and aeroelastic problems associated with flight of high speed aircraft**

Max TF Term Weight:

Document Id	**Relevant (Yes/No)**	**Reason for Irrelevant**
12	Yes	
51	Yes	
746	Yes	
1263	No	Irrelevant material based on its content Ranked high due to lemmas like “turbulent”, “transfer” etc.
884	No	Irrelevant material based on its content Ranked high due to lemmas like “structural”, “aircraft” etc.

Okapi Weight:

Document Id	**Relevant (Yes/No)**	**Reason for Irrelevant**
12	Yes	
51	Yes	
746	Yes	
1089	No	Irrelevant material based on its content Ranked high due to lemmas like “aerodynamic” etc.
172	Yes	

- **Query 3: what problems of heat conduction in composite slabs have been solved so far**

Max TF Term Weight:

Document Id	Relevant (Yes/No)	Reason for Irrelevant
485	Yes	
5	No	Ranked high due to “composite”, “slabs” etc.
399	Yes	
144	Yes	
181	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	Reason for Irrelevant
485	Yes	
5	No	Ranked high due to “composite”, “slabs” etc.
181	Yes	
399	Yes	
144	Yes	

Query 4: 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

Max TF Term Weight:

Document Id	Relevant (Yes/No)	Reason for Irrelevant
166	Yes	
488	Yes	
1061	No	Ranked high due to many lemmas common between the document and query.
1189	No	Ranked high due to many lemmas common between the document and query.
1315	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	Reason for Irrelevant
166	Yes	
1255	No	Ranked high due to many lemmas common between the document and query.
1085	No	Ranked high due to many lemmas common between the document and query like numerical etc
1189	No	Ranked high due to many lemmas common between the document and query.
1315	No	Ranked high due to many lemmas common between the document and query.

Query 5: what chemical kinetic system is applicable to hypersonic aerodynamic problems

Max TF Term Weight:

Document Id	Relevant (Yes/No)	Reason for Irrelevant
103	Yes	
1032	Yes	
943	No	Ranked high due to many lemmas common between the document and query.
625	Yes	
1272	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	
1032	Yes	
625	Yes	
103	Yes	
1272	No	Ranked high due to many lemmas common between the document and query.

943	No	Ranked high due to many lemmas common between the document and query.

Query 6: what theoretical and experimental guides do we have as to turbulent couette flow behaviour

Max TF Term Weight:

Document Id	**Relevant (Yes/No)**	
491	Yes	
386	No	Ranked high due to many lemmas common between the document and query.
257	Yes	
385	No	Ranked high due to many lemmas common between the document and query.
798	Yes	

Okapi Weight:

Document Id	**Relevant (Yes/No)**	
121	Yes	
491	Yes	
386	No	Ranked high due to many lemmas common between the document and query.
959	Yes	
610	No	Ranked high due to many lemmas common between the document and query.

Query 7: 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

Max TF Term Weight:

Document Id	**Relevant (Yes/No)**	
492	Yes	
56	Yes	

434	No	Ranked high due to many lemmas common between the document and query.
57	Yes	
124	Yes	

<u>Okapi Weight:</u>

Document Id	Relevant (Yes/No)	
492	Yes	
122	Yes	
232	Yes	
248	Yes	
57	No	Ranked high due to many lemmas common between the document and query.

<u>Query 8</u>: what methods -dash exact or approximate -dash are presently available for predicting body pressures at angle of attack

<u>Max TF Term Weight:</u>

Document Id	Relevant (Yes/No)	
122	Yes	
69	No	Ranked high due to many lemmas common between the document and query.
492	Yes	
248	Yes	
232	Yes	

<u>Okapi Weight:</u>

Document Id	Relevant (Yes/No)	
492	Yes	
122	Yes	
232	Yes	
248	Yes	

69	No	Ranked high due to many lemmas common between the document and query.

Query 9: papers on internal /slip flow/ heat transfer studies

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
550	Yes	
306	Yes	
21	Yes	
22	No	Ranked high due to many lemmas common between the document and query.
571	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
550	Yes	
21	Yes	
45	No	Ranked high due to many lemmas common between the document and query.
270	Yes	
22	No	Ranked high due to many lemmas common between the document and query.

- **Query 10: are real-gas transport properties for air available over a wide range of enthalpies and densities**

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
493	Yes	
302	Yes	
949	No	Ranked high due to many lemmas common between the document and query.

1143	Yes	
1009	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
302	Yes	
493	Yes	
1010	Yes	
524	Yes	
1264	No	Ranked high due to many lemmas common between the document and query.

Query 11: is it possible to find an analytical, similar solution of the strong blast wave problem in the newtonian approximation

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
495	Yes	
72	Yes	
654	Yes	
1327	No	Ranked high due to many lemmas common between the document and query.
1157	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
495	Yes	
472	Yes	
72	Yes	
654	Yes	
1327	No	Ranked high due to many lemmas common between the document and query.

Query 12: how can the aerodynamic performance of channel flow ground effect machines be calculated

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
624	Yes	
650	Yes	
506	Yes	
966	No	Ranked high due to many lemmas common between the document and query.
941	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
624	Yes	
650	Yes	
506	Yes	
1232	No	Ranked high due to many lemmas common between the document and query.
36	Yes	

Query 13: what is the basic mechanism of the transonic aileron buzz

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
496	Yes	
903	Yes	
502	No	Ranked high due to many lemmas common between the document and query.
38	Yes	
313	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	
496	Yes	
903	Yes	
520	Yes	
313	No	Ranked high due to many lemmas common between the document and query.
38	Yes	

Query 14: papers on shock-sound wave interaction

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
64	Yes	
291	Yes	
256	No	Ranked high due to many lemmas common between the document and query.
65	Yes	
335	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	
64	Yes	
291	Yes	
265	Yes	
256	No	Ranked high due to many lemmas common between the document and query.
568	Yes	

Query 15: material properties of photoelastic materials

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
462	Yes	
1099	Yes	
1025	Yes	
463	No	Ranked high due to many lemmas common between the document and query.
1043	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
462	Yes	
1099	Yes	
817	Yes	
463	No	Ranked high due to many lemmas common between the document and query.
1027	Yes	

- **Query 16: can the transverse potential flow about a body of revolution be calculated efficiently by an electronic computer**

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
498	Yes	
106	No	Ranked high due to many lemmas common between the document and query.
1006	Yes	
1043	Yes	
93	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
498	Yes	
106	No	Ranked high due to many lemmas common between the document and query.
1255	Yes	
231	Yes	
1301	No	Ranked high due to many lemmas common between the document and query.

- **Query 17: can the three-dimensional problem of a transverse potential flow about a body of revolution be reduced to a two-dimensional problem**

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
1108	Yes	
336	Yes	
106	Yes	
1301	No	Ranked high due to many lemmas common between the document and query.
700	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
445	Yes	
336	Yes	
1108	Yes	
801	Yes	
1301	No	Ranked high due to many lemmas common between the document and query.

Query 18: are experimental pressure distributions on bodies of revolution at angle of attack available

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
248	Yes	
197	Yes	
498	Yes	
56	No	Ranked high due to many lemmas common between the document and query.
234	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	
498	Yes	
197	Yes	
1006	No	Ranked high due to many lemmas common between the document and query.
492	Yes	
248	Yes	

Query 19: does there exist a good basic treatment of the dynamics of re-entry combining consideration of realistic effects with relative simplicity of results

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
82	Yes	
482	Yes	
1119	Yes	
554	No	Ranked high due to many lemmas common between the document and query.
706	Yes	

Okapi Weight:

Document Id	Relevant (Yes/No)	
482	Yes	
82	Yes	
1346	Yes	
274	No	Ranked high due to many lemmas common between the document and query.
706	Yes	

Query 20: has anyone formally determined the influence of joule heating, produced by the induced current, in magnetohydrodynamic free convection flows under general conditions

Max TF Term Weight:

Document Id	Relevant (Yes/No)	
500	Yes	
268	Yes	
88	Yes	
270	No	Ranked high due to many lemmas common between the document and query.
450	No	Ranked high due to many lemmas common between the document and query.

Okapi Weight:

Document Id	Relevant (Yes/No)	
500	Yes	
268	Yes	
88	Yes	
270	No	Ranked high due to many lemmas common between the document and query.
993	Yes	

Based on observations – Okapi weighting yields comparatively better results than the Max TF Term weighting schemes. The reason according to me for this is because the Okapi weighting scheme allocates weights better than max tf term for the frequent terms as it takes the document frequency and average document length into account and therefore minimizes their effect during cosine similarity calculation. The terms that are less frequent get more weight.

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

  For queries I considered the document frequency = 1 and also the collection size = 1. The reason for this is that the number of lemmas in queries is very small. If we consider the calculated values for queries like in documents then the weights come out to be negative due to the use of log in the weighting schemes.

  However, since the document collection is large – I used the collection size and document frequency as by the definition.