NIKITA VISPUTE
NET ID: NXV170005
UTD Email: nxv170005@utdallas.edu
CS 6322.001

HOMEWORK 3

PROGRAM DESCRIPTION

Resources Used:

Wordnet Lemmatizer, Stopwords, Averaged_perceptron_tagger from nltk library.

Files: HW3.py, Cranfield directory and hw3.queries

Python version used: 2.7.5(UTD Server), 3.7.1 (personal machine)

1. Turn in the vector representation of the query, and 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.

Ans: Please see HW3.txt file

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

Ans: Please see HW3.txt file

3. Identify which documents you think are relevant and non-relevant for each query.

Ans:

Query1: what similarity laws must be obeyed when constructing aeroelastic models of heated high-speed aircraft
Relevant Documents: 817, 885, 327, 180
Non-Relevant Documents: 359, 884

Query2: what are the structural and aeroelastic problems associated with flight of high-speed aircraft
Relevant Documents: 884, 875, 885
Non-Relevant Documents: 883, 1168, 320, 184

Query3: what problems of heat conduction in composite slabs have been solved so far
Relevant Documents: 485, 282, 5
Non-Relevant Documents: 181, 182, 90, 320

Query4: 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
Relevant Documents: 31, 1030, 854
Non-Relevant Documents: 181, 880, 1023, 1026

Query5: what chemical kinetic system is applicable to hypersonic aerodynamic problems
Relevant Documents: 1103, 138, 139, 140, 1102
Non-Relevant Documents: None

Query6: what theoretical and experimental guides do we have as to turbulent couette flow behaviour
Relevant Documents: 272, 776, 777
Non-Relevant Documents: 778, 1174, 883, 271

Query7: 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
Relevant Documents: 233, 759
Non-Relevant Documents: 1286, 492, 709, 137, 708, 143

Query8: what methods -dash exact or approximate -dash are presently available for predicting body pressures at angle of attack
Relevant Documents: 1083, 1286
Non-Relevant Documents: 1379, 608, 711, 745, 1287, 107

Query9: papers on internal /slip flow/ heat transfer studies \
Relevant Documents: 763, 838, 876
Non-Relevant Documents: 846, 1071, 1152, 875

Query10: are real-gas transport properties for air available over a wide range of enthalpies and densities
Relevant Documents: 1012, 1013, 1014, 503, 504
Non-Relevant Documents: None

Query11: is it possible to find an analytical, similar solution of the strong blast wave problem in the Newtonian approximation
Relevant Documents: 282, 283, 483, 484
Non-Relevant Documents: 485, 1152

Query12: how can the aerodynamic performance of channel flow ground effect machines be calculated
Relevant Documents: 631, 137, 215
Non-Relevant Documents: 543, 878, 742, 877

Query13: what is the basic mechanism of the transonic aileron buzz
Relevant Documents: 497, 644, 645
Non-Relevant Documents: 496, 643, 646, 647

Query14: papers on shock-sound wave interaction
Relevant Documents: 969, 875, 973, 291
Non-Relevant Documents: 879, 570, 256

Query15: material properties of photoelastic materials
Relevant Documents: 761, 406, 407
Non-Relevant Documents: 1096, 866, 409

Query16: can the transverse potential flow about a body of revolution be calculated efficiently by an electronic computer
Relevant Documents: 746
Non-Relevant Documents: 1359, 1359 878, 111

Query17: can the three-dimensional problem of a transverse potential flow about a body of revolution be reduced to a two-dimensional problem
Relevant Documents: 1358, 1360
Non-Relevant Documents: 282, 350 , 320

Query18: are experimental pressure distributions on bodies of revolution at angle of attack available
Relevant Documents: 803, 708
Non-Relevant Documents: 1287, 709, 254, 906

Query19: does there exist a good basic treatment of the dynamics of re-entry combining consideration of realistic effects with relative simplicity of results
Relevant Documents: 708, 224, 854
Non-Relevant Documents: 144, 709, 223, 842

Query20: has anyone formally determined the influence of joule heating, produced by the induced current, in magnetohydrodynamic free convection flows under general conditions
Relevant Documents: 500, 456
Non-Relevant Documents: 268, 270, 88, 123

4. Describe why the top-ranked non-relevant document for each query did not get a lower score.

Ans:

Ideally, all the relevant documents should have greater positive scores, and all the non-relevant documents should have scores below threshold. But in this case, the higher ranked non-relevant documents had very few query terms within the document which had more weight value. Also, the non-relevant documents did not get low scores due to their higher frequency within the document. Contrast to this, even the higher-ranked documents are not used in proper manner. The reason is, irrespective of its actual meaning, the occurrence of term within query and document is considered for relevance calculation. Thus, this cause the document to become non-relevant.

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.

Ans:

W1: Uses a form of Maximum Term Frequency
W2: Uses a form of OKAPI Term Weighting which is used for Average Document length and Document length

Weighting Scheme - W1:
It is based on the concept of Maximum Term Frequency. In this, the weights and score are calculated on the basis of the frequency of the term which are occurring in the document. The drawback of this weighting scheme is that it only looks for term's maximum occurrence and not the actual meaning of the query or the document term. This would cause it to include even the non-relevant documents due the term's occurrence irrespective of its meaning.

Weighting Scheme - W2:
It is based on the concept of length of the document. This weighting scheme neither considers the meaning of the query or document term nor it considers relevance factor. In this, if the term is present in multiple documents, it is given some negative weights, and this would cause it to include non-relevant documents which might not even has the term.
The weighting scheme W1 is preferred here as it at least considers the occurrence of terms in the query or the document.

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

The pre-processing of Cranfield documents and Query file is done after removing irrelevant terms, special characters and tags and further lemmatizing the files.
Weights are calculated for both the Document terms and query terms using both the weighing schemes W1 and W2.

W1 = (0.4 + 0.6 * log (tf + 0.5) / log (maxtf + 1.0)) * (log (collectionsize / df)/ log (collectionsize))
W2 = (0.4 + 0.6 * (tf / (tf + 0.5 + 1.5 * doclen / avgdoclen))) * log (collectionsize / df)/log (collectionsize))

The weights of each document and query terms are then normalized and later the normalized value of query terms and document terms are multiplied and then added to get the Cosine Similarity score for a particular (Query(N),Document(ID)) combination.
Thus, the ranking decision in this program is based on the Cosine Similarity score of the (Query(N),Document(ID)) combination.
The program outputs the top 5 ranked documents for every Queries under both the weighting scheme.
The program also outputs the Vector representation of the Query terms as well as the first 5 ranked document terms.