**Assignment no 4**

**Written Assignment**

**IR CSL 436 Winter 2024**

**Deadline: 25th October 2024 (5 pm)**

1. Corpus C consists of the following 3 documents:

“new york times”

“new york post”

“los angeles times”

1. Assume the term frequencies are normalized by the maximum frequency in a given document, calculate the TF-IDF weighted term vectors for all documents in C. Order the words in the vectors alphabetically.
2. Given the following query:

“new new times”

Calculate the TF-IDF weighted query vector, and compute the score of each document in C using the cosine similarity measure. Assume that term frequencies are normalized by the maximum frequency in a given query.

1. Omar has implemented a relevance feedback web search system, where he is going

to do relevance feedback based only on words in the title text returned for a page (for

efficiency). The user is going to rank 3 results. The first user, Jinxing, queries for:

banana slug

and the top three titles returned are:

banana slug Ariolimax columbianus

Santa Cruz mountains banana slug

Santa Cruz Campus Mascot

Jinxing judges the first two documents Relevant, and the third Not Relevant. Assume

that Omar’s search engine uses term frequency but no length normalization nor IDF.

Assume that he is using the Rocchio relevance feedback mechanism, with α = β =

γ = 1.

Show the final revised query that would be run.

(list the vector elements in alphabetical order).

1. Suppose that we have a standard IR evaluation data set containing 1000 documents. Assume that a particular query in this data set is deemed to be relevant to the following 25 documents in the collection:

**REL = { d1, d5, d6, d10, d88, d150, d200, d210, d250, d300, d400, d405, d450, d472, d500, d501, d530, d545, d590, d600, d635, d700, d720, d800, d900 }**

Two different retrieval systems S1 and S2 are used to retrieve ranked lists of documents from this collection using the above query. The top 10 retrieved documents for these two systems are given below (each list is in decreasing order of relevance).

RET(S1) = d2, d5, d150, d250, d11, d33, d50, d600, d500, d520  
RET(S2) = d250, d400, d150, d210, d999, d3, d501, d800, d205, d300

* 1. Compute the Precision and the Recall graphs for each system as a function of the number of documents returned (for 1 document returned, 2 documents returned, etc). First, compile your calculations in a table which shows the Precision and Recall for S1 and S2 using these query results as a function of the number of documents returned (from 1 to 10). Then, create graphs comparing Precision and Recall of the two systems.
  2. A single metric that can be used to combine precision and recall is the *F Measure.* Using the F1 meas ure (F measure with B=1), create graph similar to the above comparing the two systems.
  3. Which system is better? Explain your answer.

1. Consider the following document-term table with 10 documents and 8 terms (A through H) containing raw term frequencies. We also have a specified query, Q, with the indicated raw term weights (the bottom row in the table). Answer the following questions, and in each case give the formulas you used to perform the necessary computations.
   1. Compute the ranking score for each document based on each of the following query-document similarity measures (sort the documents in the decreasing order of the rank score):
      1. Cosine similarity
      2. Dice's coefficient (4)

A B C D E F G H

-----------------------------------------------

DOC1 0 3 4 0 0 2 4 0

DOC2 5 5 0 0 4 0 4 3

DOC3 3 0 4 3 4 0 0 5

DOC4 0 7 0 3 2 0 4 3

DOC5 0 1 0 0 0 5 4 2

DOC6 2 0 2 0 0 4 0 1

DOC7 3 5 3 4 0 0 4 2

DOC8 0 3 0 0 0 4 4 2

DOC9 0 0 3 3 3 0 0 1

DOC10 0 5 0 0 0 4 4 2

----------------------------------------------

Query 2 1 1 0 2 0 3 0

* 1. Compare the ranking obtained when, instead, binary term weights are used to the ranking obtained in part a where raw term weights were used (do this only with dot product as the similarity measure). Explain any discrepancy between the two rankings. (2)

Construct a similar table to above, but instead of raw term frequencies compute the (non-normalized) ***tf*x*idf***weights for the terms. Then compute the ranking scores using the **Cosine similarity**. Explain any significant differences between the ranking you obtained here and the Cosine ranking from the previous part.

1. Suppose that after receiving the results of a query**Q0 = "dog race"**, a user has provided relevance feedback by rating the following 3 document as **non-relevant**:

**DOC1: "greyhound race track betting"  
DOC2: "dog race betting"**   
**DOC3: "greyhound dog training"**

and the following 4documents as **relevant**:

**DOC4: "iditarod dog sled race"  
DOC5: "husky dog sled race malamute dog sled"  
DOC6: "betting alaska dog sled race"  
DOC7: "dog race alaska iditarod"**

Assuming simple term frequency weights, use Rocchio’s relevance feedback method to compute a new query **Q1** (use a positive feedback factor of 1.0 and negative feedback factor of 0.5). Show Q1 as a vector over the above index terms with the corresponding weights generated by Rocchio. Explain any significant increase or decrease in term weights. Show your work.

1. Assume that an IR system returns a ranked list of 10 total documents for a given query. Assume that according to a gold-standard labelling there are 5 relevant documents for this query, and that the only relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th positions in the ranked results. Calculate and clearly show the interpolated precision value for each of the following standard recall levels: {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} for this individual query.
2. Consider the following hypothetical information retrieval scenario. Suppose it has been found at a firm A that due to equipment malfunction, the results of blood tests taken on a particular date dd-mm-yyyy are unreliable for diabetic patients. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients’ medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

(a) Calculate the precision and recall for this system, showing the details of your calculations.

(b) Based on your results from (a), explain what the two measures mean for this scenario. How well would you say that the hospital’s information IR system works?

(c) According to the precision-recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall?

(d) For the given scenario, which measure do you think is more important, precision or recall? Why? Given your answer, what value would you give to the weighting factor α when calculating the F-score measure for the hospital’s IR system?

1. Below is a table showing how two human judges rated the relevance of a set of documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you’ve written an IR system that for this query returns the set of documents {2, 5, 6, 7, 8} and assume the documents are ranked (document 2 is 1st document in results, document 5 is 2nd document, etc.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Doc ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Judge 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| Judge 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |

1. Calculate the precision and recall of your system if a document is considered relevant only if the two judges agree it is relevant.
2. Calculate the precision and recall of your system if a document is considered relevant if either judge thinks it is relevant.
3. Calculate the average precision of your IR system based on both relevance scenarios in (a) and (b)
4. Study what is Kappa measure. Calculate the kappa measure between the two judges.
5. Is the kappa measure of acceptable value? Can the judgements be used?
6. An information retrieval system returns the following ranked list for a particular query:

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

Coloured blocks represent relevant documents; white blocks represent irrelevant documents. From the known relevance judgments, you know that there are eight relevant documents in total.

1. What is the Mean Average Precision (MAP)?
2. What is the R-precision?
3. What is Precision at 10?
4. Plot the precision-recall curve, both uninterpolated and interpolated versions.
5. Suppose you wish to find economic reports regarding the impact of oil extraction in the North Sea on the Scottish economy. A commercial document retrieval service offers the following suggested matches: the table shows how often some key phrases appear in each report.

North Sea oil Scotland economy

Report A 12, 0, 3, 24

Report B 10, 5, 20, 10

Report C 0, 12, 9, 8

Query 1, 1, 1, 1

Actually obtaining the reports will cost real money, so you would like to select the one most likely to be relevant. Your task now is to assess this using the cosine similarity measure.

(a) Write out the general formula for calculating the cosine of the angle between two 4-dimensional vectors (x1, x2, x3, x4) and (y1, y2, y3, y4).

(b) Use this formula to rank the three documents in order of relevance to the query according to the cosine similarity measure. What do you think of the results?