WSM Assignment 1

IR Evaluation Measures and Vector Space Models

---

1 . [15 points] **Evaluation Measures**

Suppose a query has a total of 10 relevant documents in a collection with 100 documents. A system has retrieved 8 documents whose relevance status is `[+,-,+,+,-,-,-,+]` in the order of ranking. A `"+"` (or `"-"`) indicates that the corresponding document is relevant (or non-relevant). For example, the first document is relevant, while the second is non-relevant, etc.

- [10/15 points] Compute the precision, recall, fallout, F1, and average precision for this result.

  $$Precision = \frac{relevant\ items\ retrieved}{retrieved\ items} = P(relevant|retrieved)$$

  We known the system has retrieved 8 documents, which has 4 relevant documents and 4 non-relevant documents. So the precision of this query result will be $\frac{4}{8}$, also as $\frac{1}{2}$.

  $$Recall = \frac{relevant\ items\ retrieved}{relevant\ items} = P(retrieved|relevant)$$

  Since we known there are 10 relevant documents in total, so the recall will be $\frac{4}{10}$, also as $\frac{2}{5}$.

  $$Fallout = \frac{not\ relevant\ items\ retrieved}{not\ relevant\ items}$$

  I found this equation from https://sites.pitt.edu/~peterb/2140-051/L4.pdf , page 12. The fallout will be $\frac{4}{95}$.

  $$F1 = 2\,\frac{precision * recall}{precision + recall} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

  F1-score, also known as balanced F measure. By using the results of previous questions, we may know the precision is $\frac{1}{2}$, and recall is $\frac{2}{5}$ to compute the F1 score.

  $$F1 = 2 * \frac{\frac{1}{2} * \frac{2}{5}}{\frac{1}{2} + \frac{2}{5}}$$

$$= 2 * \frac{\frac{1}{5}}{\frac{9}{10}}$$

$$= \frac{4}{9}$$

$$AveragePrecision = \frac{1}{|R|} \sum_{k=1}^{|R|} Precision(R_k)$$

$|R|$: number of relevant itemsR_k : k th relevant's item

by not considering the noise(without interpolated)

$$AP = \frac{1}{4}(\frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{8})$$

$$= \frac{35}{48}$$

- [5/15 points] What is the precision at the recall level of 0.15 according to 11-point interpolated?

  Assume retrieved n [+, -, +, +, -, -, -, +] documents, for each k belongs to [0, n], examine the set of top k retrieved documents.

  recall: [1/10, 1/10, 2/10, 3/10, 3/10, 3/10, 3/10, 4/10]

  precision: [1/1, 1/2, 2/3, 3/4, 3/5, 3/6, 3/7, 4/8]

$$P_{11-pt} = \frac{1}{11} \sum_{j=0}^{10} \frac{1}{N} \sum_{i=1}^{N} \tilde{P_i}(r_j)$$

$\therefore 0.7 < Precision < 0.8$ (take maximum of next future point)

---

2 . [20 points] **Evaluation Measures**

Consider the following two systems A and B with their corresponding search results, in which there are 5 relevance grades (0,1: non-relevant, 2, 3,4,5: relevant). Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

`SystemA`: 0 3 2 0 5 0 0 0 0 4

`SystemB`: 4 0 0 0 5 0 2 0 0 3

- [5/20 points] What is the MAP of each system?

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

$$MAP(SystemA) = \frac{1}{4}(\frac{1}{2} + \frac{2}{3} + \frac{3}{5} + \frac{4}{10})$$

$$= \frac{13}{24}$$

$$MAP(SystemB) = \frac{1}{4}(1 + \frac{2}{5} + \frac{3}{7} + \frac{4}{10})$$

$$= \frac{39}{70}$$

- [5/20 points] What is the R-precision of each system? Does it rank the systems the same as MAP?

$$R\text{-}precision(SystemA) = \frac{2}{4} = \frac{1}{2}$$

$$R\text{-}precision(SystemB) = \frac{1}{4}$$

$\because MAP(SystemA)$ is greater than $MAP(SystemB)$, so the $R\text{-}precision(SystemA)$ and $R\text{-}precision(SystemB)$,

$\therefore$ we can say the R-precision will rank the systems the same as MAP.

- [5/20 points] What is the NDCG of each system?

$$CumulativeGain[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i-1] + G[i], & \text{otherwise} \end{cases}$$

$$DiscountedCumulativeGain[i] = \begin{cases} G[i], & \text{if } i = 1 \\ DCG[i-1] + G[i]/\log_2 i, & \text{otherwise} \end{cases}$$

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj}\sum_{m=1}^{k}\frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

System A:

$$\text{ranking} :< 0, 3, 2, 0, 5, 0, 0, 0, 0, 4 >$$

$$CG' =< 0, 3, 5, 5, 10, 10, 10, 10, 10, 14 >$$

$$NDCG(SystemA) = 0.54$$

System B:

$$\text{ranking} :< 4, 0, 0, 0, 5, 0, 2, 0, 0, 3 >$$

$$CG' =< 4, 4, 4, 4, 9, 9, 11, 11, 11, 14 >$$

$$NDCG(SystemB) = 0.59$$

- [5/20 points] What is the main difference between MAP and NDCG?

  The main difference between MAP and NDCG will be MAP is kinda showing the distribution of the retrieved result, and MAP care only binary (which relevant=1 or non-relevant=0). The NDCG care about how the documents is ranked for different level of relevant (documents can be ranked as 0/1 for non-relevant, 2/3 for relevant, 4/5 for highly relevant and so on).

---

3 . [20 points] **Evaluation Measures**

The following list of Rs and Ns represents relevant (R) and non-relevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

`N N R N N  N R N N R  R N N N N  R N N R N`

- [5/20 points] What is the precision of the system on the top 10?

  top 10: [N, N, R, N, N, N, R, N, N, R]

  relevant items retrieved: 3, retrieved items: 10

$$\therefore Precision = \frac{3}{10}$$

- [5/20 points] What is the F1 on the top 10?

  top 10: [N, N, R, N, N, N, R, N, N, R]

  the recall on the top 10 will be $\frac{\text{relevant items retrived}}{\text{relevant items}}$, is $\frac{3}{8}$

$$\therefore F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

$$= 2\frac{\frac{3}{10} * \frac{3}{8}}{\frac{3}{10} + \frac{3}{8}}$$

$$= \frac{1}{3}$$

- [5/20 points] What is the un-interpolated precision of the system at 25% recall?

  there are 8 relevant items in the collection,

  recall: [0/8, 0/8, 1/8, 1/8, 1/8, 1/8, 2/8, 2/8, 2/8, 3/8, 4/8, 4/8, 4/8, 4/8, 4/8, 5/8, 5/8, 5/8, 6/8, 6/8]

  precision: [0/1, 0/2, 1/3, 1/4, 1/5, 1/6, 2/7, 2/8, 2/9, 3/10, 4/11, 4/12, 4/13, 4/14, 4/15, 5/16, 5/17, 5/18, 6/19, 6/20]

$$0.2 < Precision < 0.25$$

- [5/20 points] What is the interpolated precision at 33% recall?

$$interpolated\text{-}Precision = 0.35$$

---

4 . [10 points] **Vector Space Model**

Consider a collection with the documents:

`Document1`: vegetables are very healthy

`Document2`: healthy diets are very very very important

- [5/10] What would be the tf-idf vector of each document?

  IDF of word (w) is defined as: from https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d

$$IDF(w, D) = \ln(\frac{\text{total number of documents (N) in corpus D}}{\text{number of documents containing w}})$$

	vegetables	are	very	healthy	diets	important
TF-Document 1	1/4	1/4	1/4	1/4	0	0
TF-Document 2	0	1/7	3/7	1/7	1/7	1/7
IDF	2/1	2/2	2/2	2/2	2/1	2/1

- [5/10] What would be the similarity score for the documents using cosine similarity?

  by not using tf-idf vector:

$$x = (1/4, 1/4, 1/4, 1/4, 0, 0)$$

$$y = (0, 1/7, 3/7, 1/7, 1/7, 1/7)$$

$$CosSimilarity(x, y) = \frac{xy}{||x||\,||y||}$$

$$= \frac{1/4 * 0 + 1/4 * 1/7 + 1/4 * 3/7 + 1/4 * 1/7 + 0 * 1/7 + 0 * 1/7}{\sqrt{(1/4)^2 + (1/4)^2 + (1/4)^2 + (1/4)^2} * \sqrt{(1/7)^2 + (3/7)^2 + (1/7)^2 + (1/7)^2 + (1/7)^2}}$$

$$= \frac{5}{2\sqrt{13}}$$

---

5 . [10 points] **Vector Space Model**

Consider a collection with the documents and the given query:

`Document1`: federer teams up with Nadal for doubles match

`Document2`: federer is the king of Wimbledon Championships because he won eight times champion

`Query`: federer won champion

- [5/10] Assume there are no stop words. What would be the tf vector of each document?

	TF-document 1	TF-document 2
federer	1/8	1/13
teams	1/8	0
up	1/8	0
with	1/8	0

	TF-document 1	TF-document 2
Nadal	1/8	0
for	1/8	0
doubles	1/8	0
match	1/8	0
is	0	1/13
the	0	1/13
king	0	1/13
of	0	1/13
Wimbledon	0	1/13
Championships	0	1/13
because	0	1/13
he	0	1/13
won	0	1/13
eight	0	1/13
times	0	1/13
champion	0	1/13

- [5/10] Assume the stop words are: the, of, to, is, and, because, in, has, not. What would be the similarity score of the query with each document using Jaccard coefficient?

	TF-document 1	TF-document 2
federer	1/8	1/9
teams	1/8	0
up	1/8	0
with	1/8	0
Nadal	1/8	0
for	1/8	0
doubles	1/8	0
match	1/8	0
king	0	1/9
Wimbledon	0	1/9
Championships	0	1/9
he	0	1/9
won	0	1/9
eight	0	1/9
times	0	1/9
champion	0	1/9

---

6 . [25 points] **Vector Space Model**.

Consider the following table of term frequencies and document frequencies for 3 documents, which are denoted as `Doc1`, `Doc2`, `Doc3`, and suppose that there are 70,000 documents in total within the document collection.

Term	Doc1	Doc2	Doc3	
car	28	5	24	18165
auto	6	30	0	6723
insurance	0	30	30	19241
best	15	0	19	25235

- [5/25] Compute the tf-idf weights for the terms "car", "auto", "insurance", "best", for each document.

$$tf = \frac{\text{number of times the term appears in document}}{\text{total number of terms in the document}}$$

$$idf = \log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} * \text{idf}_t$$

- [5/25] Can the tf-idf weight of a term in a document exceed 1? Why or why not?

  Yes. We wish to score a term with respect to a relevant document by using tf-idf. There will be naive case the tf-idf can be 1, whenever we do log in our equation.

- [5/25] Compute the Euclidean normalized document vector for representing each documents, in each of which vector there are the four terms.

- [10/25] Rank the three documents with respect to the query "best car insurance" according to the following settings of query term weighting:

  - The weight of a term is 1 if present in the query, 0 otherwise.
  - query tf * normalized idf of the collection.