Natural Language Processing Assignment 2

Yajneesh Gowtham CS19B009
Chaitra ME19B092

March 20 2023

Question 1

Term	Document
herbivores	S1
typically	S1, S2
plant	S1, S2
eaters	S1, S2
meat	S1, S2
carnivores	S2
deers	S3
eat	S3
grass	S3
leaves	S3

Question 2

The formula used to calculate is as follows:
TF = number of times the term occurs in document
$IDF = log(N/df)$
where
N = number of documents
df = number of documents containing the term
$TF - IDF = TF \cdot IDF$

The TF-IDF vectors for S1, S2 and S3 are as follows:

$$S1 : (0.477, 0.176, 0.176, 0.352, 0.176, 0, 0, 0, 0)$$

$$S2 : (0, 0.176, 0.176, 0.352, 0.176, 0.477, 0, 0, 0, 0)$$

$$S3 : (0, 0, 0, 0, 0, 0, 0.477, 0.477, 0.477, 0.477)$$

Question 3

Based on the inverted index, S1 and S2 would be retrieved for the query "plant eaters" since these documents contain both the terms "plant" and "eater". S3 would not be retrieved since it does not contain "plant" or "eater".

S1: "Herbivores are typically plant eaters and not meat eaters"
S2: "Carnivores are typically meat eaters and not plant eaters"

Question 4

The TF-IDF vector for the query matrix is as follows:
"plant eater" or

$$q : (0, 0, 0.176, 0.176, 0, 0, 0, 0, 0, 0)$$

$$\text{Length of vector } q = 0.249$$

$$\text{Length of vector } S1 = 0.666$$

$$\text{Length of vector } S2 = 0.666$$

The cosine similarity is given by:

$$cosineSim(q, S1) = (0.176 \cdot 0.176 + 0.176 \cdot 0.352) / 0.249 \cdot 0.666 = 0.559$$

$$cosineSim(q, S2) = (0.176 \cdot 0.176 + 0.176 \cdot 0.352) / 0.249 \cdot 0.666 = 0.559$$

Since both cosine similarities are equal, a tie-breaking metric is needed to rank these documents. Since both document lengths and term frequencies are equal, we would need a more context-specific measure.

Question 5

The ranking we get above may not be ideal. Since the query is "plant eaters" the most relevant document could be S3 "Deers eat grass and leaves" which implies that deers are plant eaters. S2: "Carnivores are typically meat eaters and not plant eaters" may not be relevant at all since it actually describes carnivores who are not plant eaters.

Question 7

The inverse document frequency(IDF) of a term is represented by `log(N/n)` where N is the total number of documents and n is the number of documents in which the term occured.

The idf value of a term that occurs in every term is zero since the value of n=N

$$log_2(N/n) = log_2(N/N) = log_2(1) = 0$$

The case term frequency can be zero for a query term which doesn't exist in the corpus then then we can just add 1 to the denominator to avoid zeros in the denominator to get finite values and the formula can be updated as follows

$$log(N/(n+1))$$

Question 8

The other similarity/distance measures that can be used to compare vectors are as follows

- Jaccard distance : it is the similarity of intersection between two sets divided by sum of two sets here we are relying on terms to match i.e only on TF not on IDF
- Euclidean distance : It is associated as distance between two vectors i.e L2-norm between vectors it can give a false sense of similarity because even when two vectors are distant the angle between them could be very less and thus similar
- Minkowski distance : It is associated as a Lp-norm between vectors where generally p is taken as 1 or 2 with 1 being manhattan distance and 2 being Minkowski distance the draw back is same as euclidean distance

Question 9

- Since In IR systems the data is extremely skewed that is more than 99% of the documents are not relevant so in measuring accuracy we cannot capture only the information pertaining to relevant documents
- One of the important functions of an IR system is to provide the most relevant documents first. Since accuracy does not provide information about the relative relevance of documents, that is the ranking, it is not the best metric to evaluate IR systems. Instead, other metrics such as Precision , Recall , F1 score , Mean Average Precision and nDCG are more appropriate.

Question 10

$$F_\alpha = \left(\frac{\alpha}{precision} + \frac{1-\alpha}{recall} \right)^{-1}$$

where

$$0 \leq \alpha \leq 1$$

So for recall to be given more weight than precision the value of

$$0 \leq \alpha < 0.5$$

Question 11

Precision @k measures the fraction of relevant documents out of the top k retrieved documents. It does not take into account the order of these k documents. In, contrast, Average precision

@k takes into account the ranking of the top k retrieved documents and is a more nuanced metric for evaluating the performance of the IR system.

Question 12

Average precision @k is based on the evaluation of one query whereas Mean Average precision @k is the average of M queries. The formulas are as follows:

$$Averageprecision@K = \frac{1}{r}\Sigma_{k=1}^{K} Precision@k \cdot rel(k)$$

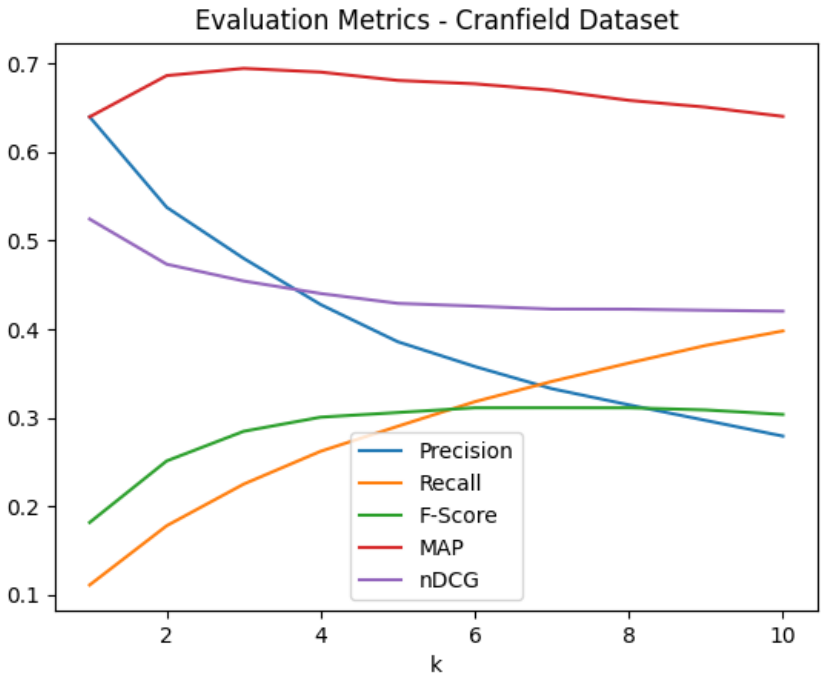
$$Meanaverageprecision@K = \frac{1}{M}\Sigma_{m=1}^{M}\frac{1}{r}\Sigma_{k=1}^{K} Precision@k \cdot rel(k)$$

Question 13

nDCG is preferred over AP in cranfield dataset because average precision only gives binary weights to the retrieved documents whereas in nDCG we give weights to each document based on relevance since cranfield dataset also has relevance values to each query.

Question 15

The output is as follows



Observations:

- As the value of K increases precision decreases which shows that relevant documents are being retrieved at first
- As the value of k increases recall increases since more and more relevant documents are retrieved
- F score which is an average of both precision and recall first increases and maintains the max value and decreases
- MAP values increases at first and then decreases continously shows that number of relevant documents are less for further retrievals
- nDCG value remains to be almost same through out although it is higher initially it infers that there are no highly relevant documents that needs to be retrieved as k increases.

Question 16

- The following queries have no relevant documents being retrieved in the top 10
- The query numbers are 9,19,22,28,35,44,62,63,74,85,87,110,115,116,117,151,152,167,207,216 since for some the relevant documents have doesn't contain some key words of the query which are present in other queries that get the advantage over other

Question 17

The shortcomings of the vector space model are as follows

- The query words must exactly match the terms of the documents or else we cannot retrieve searches for the query
- Documents with similar context but different terms won't be matched similar so some documents may not be retrieved
- The order in which the terms appear are lost in vector space representation
- Assumes that terms are independent of each other
- Long documents are poorly represented they have a very high dimensionality

Question 18

The way to include the title with the document is to concatenate the title with the document if we want to give 3 times the weight to the title then concatenate the title three times with the document

Question 19

- Advantages:

1. Bigrams capture the context of the words and meaning better than unigrams this helps to maintain some order in which the terms appear
2. Bigrams help to reduce the impact of stop words also removing stop words which might be useful
3. Bigrams help to capture the structure of sentences better than unigrams.

- Disadvantages:
  1. Bigrams increase the size of the index i.e, the number of terms worst case it will be order of N^2 where N is the number of terms in unigrams
  2. Bigrams increase the computational complexity of the search engine
  3. Bigrams can sometimes give less accurate results since now in vector space model every two words must exactly match or else similarity will not be there between the documents.

Question 20

The following are ways to use implicit feedback without user input for query -document relevance

- One way is to click through data i.e, the amount of clicks a user clicks on a page for a particular query, if he clicks more then that page is more relevant to the query else less relevant
- Another way is to take into account the user's behaviour the amount of time he spent on a page and number of times he is redirected to the page and the number of pages viewed