SECTION A

1.

(a)

The following documents have been processed by an IR system where stemming is not applied:

DocID	Text
Doc1	breakthrough vaccine for covid19
Doc2	new covid19 vaccine is approved
Doc3	new approach for treating patients
Doc4	new hopes for new covid19 patients in the world

(i) Assume that the following terms are stopwords: in, is, for, the. Construct an inverted file for these documents, showing clearly the dictionary and posting list components. Your inverted file needs to store sufficient information for computing a simple tf*idf term weight, where $w_{ij} = tf_{ij} * \log_2(N/df_i)$

[5]

(ii) Compute the term weights of the two terms "breakthrough" and "vaccine" in Doc1. Show your working.

[2]

(iii) Assuming the use of a best match ranking algorithm, rank all documents using their relevance scores for the following query:

covid19 vaccine

Show your working. Note that $\log_2(0.75) = -0.4150$ and $\log_2(1.3333) = 0.4150$.

[3]

(iv) Typically, a log scale is applied to the tf (term frequency) component when scoring documents using a simple tf*idf term weighting scheme. Explain why this is the case illustrating your answer with a suitable example in IR. Explain through examples how models such as BM25 and PL2 control the term frequency counts.

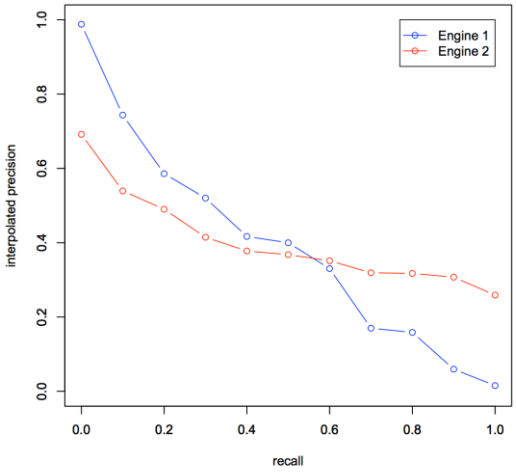
[4]

(b) Consider the recall-precision graph below showing the performances of two variants of a search engine that mimic Google Scholar on a collection of research papers. There is no difference between the two variants apart from how they score documents. Assume that you are a student looking to find all published papers on a

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

given topic. In other words, you do not want to miss any of the relevant documents. Explain which search engine will be more suitable for your task and why?

[5]



(c) Assume that you have decided to modify the approach you use to rank the documents of your collection. You have developed a new Web ranking approach that makes use of recent advances in neural networks. Explain in detail the steps you need to undertake to determine whether your new Web ranking approach produces a better retrieval performance than the original ranking approach.

Assignment Project Exam Help
https://powcoder.com

[5]

(d) Consider a query with two terms, whose posting lists are as follows:

Add WeChat powcoder

term1 → [id=2, tf=2], [id=5, tf=1], [id=6, tf=1]

term2 → [id=2, tf=4], [id=4, tf=3] , [id=5, tf=4]

Explain and provide the exact steps/order in which the posting lists will be traversed by the TAAT & DAAT query evaluation strategies and the memory requirements of both strategies for obtaining a result set of K documents from a corpus of N documents (K<N).

[6]

2.

(a) Consider a corpus of documents C written in English, where the frequency distribution of words approximately follows Zipf's law r * $p(w_r|C) = 0.1$, where r = 1,2, …, n is the rank of a word by decreasing order of frequency. Hence, the words are ordered by decreasing order of probability of occurrence in the corpus such that w_r is the word at rank r, and $p(w_r|C)$ is the probability of occurrence of word w_r in the corpus C.

What proportion of word occurrences would be removed from the collection if we ignored all occurrences of the five most frequent words in the collection? *Justify your answer.*

[5]

(b) Consider the query "jackson music" and the following term frequencies for the three documents D1, D2 and D3, where the search engine is using raw term frequency (TF) but no IDF:

	indiana	jackson	life	michael	music	pop
D1	0	4	0	3	0	6
D2	4	0	3	4	0	0
D3	0	3	0	5	4	4

Assignment Project Exam Help

Assume that the system has returned the following ranking: D2, D3, D1. The user judges D3 to be relevant and both D1 and D2 to be non-relevant.

https://powcoder.com

(i) Show the original query vector, clearly stating the dimensions of the vector.

Add WeChat powcoder

[2]

(ii) Use Rocchio's relevance feedback algorithm (with $\alpha=\beta=\gamma=1$) to provide a revised query vector for "jackson music". Terms in the revised query that have negative weights can be dropped, i.e. their weights can be changed back to 0. *Show all your calculations*.

[4]

(c) Suppose we have a corpus of documents with a dictionary of 8 words w_1, ..., w_8. Consider the table below, which provides the estimated language model p(w|C) using the entire corpus of documents C (second column) as well as the raw word counts in doc_1 (third column), where $ct(w, doc_i)$ is the raw count of word w (i.e. its *term frequency*) in document doc_i. The fourth column corresponds to a classical unigram language model for document doc_1 estimated using the non-smoothed maximum likelihood estimator.

Word	p(w|C)	$ct(w, doc_1)$	$p_{lm}(w, doc_1)$
w_1	0.4	2	0.2
w_2	0.15	2	

w_3	0.05	1	
w_4	0.1	2	
w_5	0.05	2	
w_6	0.15	0	
w_7	0.05	1	
w_8	0.05	0	

(i) Provide the missing values in the table for the non-smoothed maximum likelihood probabilities $p_{lm}(w|doc_1)$ for each of the 8 words (fourth column). *Show your calculations*.

[4]

(ii) Suppose we now smooth the language model for doc_1 using the Dirichlet prior smoothing method with parameter $\mu = 10$. Recall that for a given word w, the smoothed probability using the Dirichlet prior smoothing method is estimated as follows:

$$p(w|doc_1) = \frac{ct(w|doc_1) + \mu p(w|C)}{|doc_1| + \mu}$$

Assignment Project Exam Help

where $|doc_1|$ is the document length of doc_1 in tokens.

https://powcoder.com

Compute the Dirichlet smoothed probabilities for words w_1 and w_2 in Doc_1. *Show your calculations.*

Add WeChat powcoder

[2]

(iii) For the remaining 6 words of doc_1 (w_3, w_4, w_5, w_6, w_7, w_8), explain whether the smoothed probability will be larger than, equal to, or smaller than the initial non-smoothed maximum likelihood estimate. You do not have to compute the actual probabilities, but just use one of {> , = , <} to indicate the expected change. *You must justify your answer.*

[3]

(iv) Let $q = w_1\ w_6$ be the query issued by the user. Provide the probability of q according to the Dirichlet smoothed language model for doc_1 (recall that $\mu = 10$). *Show your calculations*.

[2]

(v) Assume that we make the value of μ larger (i.e. > 10). Explain if the probability of q will become larger, smaller or if it will remain the same. *Justify your answer*.

[2]

(vi) Assume another document doc_2 in the corpus, which is identical to doc_1 with the exception that one occurrence of w_1 has been changed to word w_5. Hence, we have $ct(w_1, doc_2) = 1$ and $ct(w_5, doc_2) = 3$.

Let $q_1 = w_1\ w_5$ be the new query.

If no smoothing is applied, using the query likelihood retrieval method, state which of the two documents (doc_1 or doc_2) will be ranked higher. Justify you answer.

Using the query likelihood retrieval method but this time with Dirichlet prior smoothing applied ($\mu = 10$), show which of the two documents (doc_1 or doc_2) would be ranked higher. *Show your calculations.*

Discuss whether smoothing has an impact on the ranking order of doc_1 and doc_2 and how? *Justify your answer.*

[6]

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

SECTION B

3. (a)

Consider the following vector space scoring formula:

$$Score(d,q) = \sum_{w \in d, w \in q} ct(w,q) * ct(w,d) * \frac{N_w + 1}{M + 1}$$

where *ct(w,d)* and *ct(w, q)* are the raw counts of word *w* in document *d* and query *q*, respectively (in other words, the term frequency of *w* in *d* and *q*, respectively); N_w is the number of documents in the corpus that contain word *w*, and *M* is the total number of documents in the corpus. Provide 4 reasons why the retrieval formula above is very unlikely to perform well in a Web search context. *Justify your answers.*

[5]

Assignment Project Exam Help

(b) For a particular query q, the multi-grade relevance judgements of all documents are {(d1,1),(d3, 4),(d6, 2),(d9, 3),(d11, 1),(d31, 2)}, where each tuple represents a document ID and a relevance judgment pair, and all the other documents are judged as non-relevant. Documents are judged on the scale 0-4 (0:not relevant – 4:highly relevant). Two IR systems return their retrieval results with respect to this query as follows (these are all results they have returned for this query):

https://powcoder.com

Add WeChat powcoder

System A: {d1, d2, d3, d4, d5, d6, d7}
System B: {d31, d22, d3, d6, d15}

For both System A and System B, compute the following ranking evaluation metrics. You must clearly articulate how you compute each of these metrics. Since there are two DCG definitions discussed in the class, you should use the original one where *1/log₂ (rank)* is used as the discount factor that is applied to the gain:

(i) Average Precision (AP). *Show your calculations.*

[3]

(ii) Normalised Discounted Cumulative Gain (NDCG) for each rank position. In your answer, provide the ideal DCG values for the perfect ranking for the given query. You might wish to note that $\log_2 2 = 1$; $\log_2 3 = 1.59$; $\log_2 4 = 2$; $\log_2 5 = 2.32$; $\log_2 6 = 2.59$ and $\log_2 7 = 2.81$. *Show your calculations.*

[6]

(c) URL length has been shown to be an important feature for some Web search tasks. Discuss which types of information needs on the Web, the URL length feature is most appropriate for.

Consider a linear learning to rank model for Web search using 4 features: PL2, Proximity, PageRank and URL length. Using such a model, explain the main disadvantage of using linear learning to rank models in Web search.

[5]

(d) A posting list for a term in an inverted index contains the following three entries:

id=3 tf=4 id=7 tf=3 id=10 tf=5

Applying the delta compression of ids, show the binary form of the unary compressed posting list. What is the resulting (mean) compression rate, in bits per integer?

Assignment Project Exam Help

[5]

https://powcoder.com

(e) A Web search engine has devised a new interface to present its search results. Describe three specific approaches that could be used by the search engine to evaluate the interface change.

Add WeChat powcoder

Which approach you would recommend and why?

[6]