CS657A: Information Retrieval

End-Sem: Marks = 80

28th April, 2022, 5:00–6:45pm

Take a moment to read the instructions and the question carefully before you jump into answering it.
Answer the parts of a question in order and in separate pages.
The time limit is **1 hour and 45 minutes**. There are 6 questions on 2 pages.

Q1: [15 marks] Mark as *true* or *false*: +1 for every right answer, and -1 for every wrong answer. Put an 'X' for every statement that you skip answering.

(a) The bag-of-words model for Boolean retrieval system treats a document as a multi-set of terms. T

(b) The size of postings list is the same as document frequency. T

(c) Diacritic marks, such as "naïve" versus "naive" can always be treated as equivalent. F

(d) In the most basic tf-idf model, the tf-idf score for a term t in a document d is 0 if and only if t is absent in d. T

(e) Champion lists pre-compute the top scoring documents for each term. T

(f) All the members of the BMx family (such as BM25) uses document length in some form or other. F

(g) The main difference between inference and belief networks is the way the relationship between documents and terms are handled. T

(h) The basic idea of ranking using language models is by the probability of generating the query using the language model of the document. T

(i) In LDA, the per-topic word distribution function remains the same for all documents.

(j) GloVe can handle out-of-vocabulary words while Word2Vec cannot.

(k) Self-attention essentially means a weighted version of another unit as an additional input to the current unit. T

(l) Both FastText and BERT produce contextual vectors.

(m) BART has both an encoder and a decoder.

(n) XLNet uses permutations of sentences as input.

(o) Siamese networks constrain the weights to be the same.

Q2: Consider the documents d_1, d_2, d_3 and the query q:
d_1: Shipment of gold damaged in a fire.
d_2: Delivery of silver arrived in a silver truck.
d_3: Shipment of gold arrived in a truck.
Query q: gold silver truck

(a) [10 marks] Rank the three documents in terms of tf-idf.

(b) [10 marks] Rank the three documents in terms of cosine similarity with the query.

(c) [10 marks] Rank the three documents in terms of probabilistic odds in the Binary Independence Model. Assume that d_1 is irrelevant and d_2, d_3 are relevant.

Q3: Suppose the correct answer set for a retrieval task is $\{d_1, d_2, d_3, d_4, d_5\}$.

(a) [5 marks] Suppose a retrieval algorithm A_1 retrieve documents in the following order:
$A_1 : \{d_3, d_9, d_2, d_5, d_1\}$
Find MAP@5 of the algorithm.

(b) [5 marks] Suppose a retrieval algorithm A_2 retrieve documents in the following order:
$A_2 : \{d_3, d_8, d_4, d_7, d_1\}$
Find BPREF of the algorithm.

(c) [5 marks] Suppose a retrieval algorithm A_4 retrieve documents in the following order:
$A_4 : \{d_3, d_2, d_1, d_5, d_4\}$
Find Kendall's tau for the algorithm.

Q4: [5 marks] Suppose for a set of documents $\{d_1, d_2, d_3, d_4, d_5\}$, the documents d_1, d_2 are relevant.

Suppose a retrieval algorithm A retrieves d_1, d_3, d_4 as the relevant set.

Find precision, recall and F-score of the algorithm.

Q5: [5 marks] Consider two experts who mark the documents in a retrieved rank list as follows:

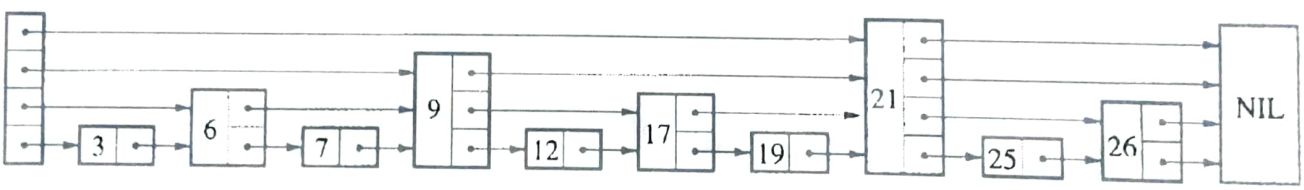
E_1 :	R	I	R	R	I	I	R	R	R	I
E_2 :	R	R	I	I	I	I	R	I	I	I

Compute Cohen's kappa between the two experts.

Q6: [10 marks] Consider a multi-tier skip list. A first-level link points to the next element; a second-level link points to the element 2 positions away; a third-level link to 4 positions away, and so on. The first entry (index 0) has a "next" pointer to index 1, index 2, index 4, etc. till 2^l.

Assume a m-length sorted linked list. The number of levels is l such that $2^l + 1$ is less than m.

An example of such a multi-tier skip list is shown below:



For a m-length list, find the average-case *search* complexity of an element.