Mid-Term Examination, Autumn 2022-23
Subject No.: CS60092 Subject: Information Retrieval
Department: Computer Science and Engineering
Total Marks: 60 Time: 2 hours

Important Instructions

1. Attempt all questions. All parts of the same question must be answered together.
2. No clarifications can be provided during the exam. Make your own reasonable assumptions if necessary. Explicitly mention any assumptions you make.
3. All workings must be shown. You can use calculators.
4. There will be partial marking. Some questions may seem open-ended; marks for these will be given based on well-grounded reasoning and ingenuity.

Best of Luck!

Problem 1

Fill in the blanks.

(a) A document collection which has 50 relevant documents, 20 of which are relevant and 23 are irrelevant; the precision of the system is _____.

(b) The most common words such as articles, prepositions etc. are removed from tokens in the step _____.

(c) Given a document containing the sentence "I left my left bag at my home" the number of tokens in the sentence is _____.

(d) A crude heuristic process that chops off the ends of the words to reduce inflectional forms of words and reduce the size of the vocabulary is called _____.

(e) The number of times that a word or term occurs in a document is called the _____.

5 marks

Problem 2

Consider the following documents collection:
d_1 = "Big cats are nice and funny"
d_2 = "Small dogs are better than big dogs"
d_3 = "Small cats are afraid of small dogs"
d_4 = "Big cats are not afraid of small dogs"
d_5 = "Funny cats are not afraid of small dogs"

(a) Compute the tokens for each document. Call this set S_1.

R	q_1	q_2
1	**A**	**F**
2	L	**G**
3	**G**	D
4	F	**E**
5	D	L
6	**E**	I
7	**B**	H
8	**H**	C
9	I	**B**
10	C	A

Table 1: Response of the IR system.

(b) Normalize the tokens with respect to plurals and upper/lower case. Call this set S_2.

(c) Compute the dictionary relative to the documents collection.

2+1+1=4 marks

Problem 3

Starting from the documents collection of **Problem 2**, build the documents-terms incidence matrix as required by the Boolean retrieval model. Assume set S_2 to be the set of all tokens. Hence compute the following queries: (a) q_1 = funny **AND** dog (b) q_2 = nice **OR** dog (c) q_3 = big **AND** dog **AND NOT** funny

3+1+1+2=7 marks

Problem 4

Starting from the documents collection of **Problem 2**, build an inverted index for the documents collection. Assume set S_2 to be the set of all tokens.

5 marks

Problem 5

Consider a document collection made of 100 documents. Given a query q, the set of documents relevant to the users is $D = \{d_3, d_{12}, d_{34}, d_{56}, d_{98}\}$. An IR system retrieves the following documents $D = \{d_3, d_{12}, d_{35}, d_{56}, d_{66}, d_{88}, d_{95}\}$.

(a) Compute the number of true-positives, true-negatives, false-positives, and false-negatives.

(b) Compute the precision and the recall.

2+2=4 marks

Problem 6

An IR system produces the rankings shown in Table 1 in answer to queries q_1 and q_2. The bold and underscored documents are the ones relevant to the user. Compute the mean average precision of the system.

5 marks

Problem 7

Consider a corpus of documents $\mathcal{S}$, having the following documents $D_1, D_2, \ldots, D_5$ (see Table 2):

Documents	Words
D_1	Data Base System Concepts
D_2	Introduction to Algorithms
D_3	Computational Geometry: Algorithms and Applications
D_4	Data Structures and Algorithm Analysis on Massive Data Sets
D_5	Computer Organization

Table 2: Corpus for Question 7

Due to stemming, "computer" and "computational" are treated as the same term. The same applies for "algorithms" and "algorithm". Hence, after pre-processing, we get the following vocabulary of 8 words: $\{w_1$ = data, w_2 = system, w_3 = algorithm, w_4 = computer, w_5 = geometry, w_6 = structure, w_7 = analysis, w_8 = organization$\}$

(a) Let $\mathrm{tf}(w, D)$ denote the term frequency of a term w in a document D. Calculate the values $\mathrm{tf}(w_i, D_j)\ \forall\ 1 \leqslant i \leqslant 8, 1 \leqslant j \leqslant 5$.

(b) Let $\mathrm{idf}(w)$ denote the inverse document frequency of a term w. Calculate $\mathrm{idf}(w_i)\ \forall\ 1 \leqslant i \leqslant 8$. [Use $\mathrm{idf}(w) = \log_2\left(\frac{N}{df(w)}\right)$]

(c) Convert each document D into a point in 8-dimensional vector space, based on the tf-idf model. Specificially, use the exact formulation $\mathrm{tfidf}(w_i, D_j) = \log_2(1 + \mathrm{tf}(w_i, D_j)) \cdot \mathrm{idf}(w_i)$.

(d) Suppose we have a query Q: "Geometry Algorithm Concepts". Convert Q to a 8-dimensional point in the same vector space using the same tf-idf formulation.

(e) Rank the documents $D_1, D_2, \ldots, D_5$ in descending order of relevance to the query Q according to the cosine similarity metric.

2+2+5+2+4=15 marks

Problem 8

Answer the following questions (short, to-the-point answers expected):

(a) Write down all the permuterm indices for the word "fluffy".

(b) What will be the permuterm keys to perform lookup on, for the wildcard queries *s*ng* and *man**?

(c) Write two points of difference between Blocked Sort-Based Indexing (BSBI) and Single-Pass In-Memory Indexing (SPIMI).

(d) How would you perform the permuterm retrieval for the wildcard query *se*n*tion*?

(e) What would be the Boolean bi-gram query equivalent to the query *se*n*tion*?

(f) Consider the document "digital camera video camera" and the query "flagship video camera video". Calculate the score between the given document and query based on "lnc.ltc" scheme. Assume that there are a total of 100,000 documents, and the document frequencies of the words "digital", "camera", "video" and "flagship" are 20000, 5000, 10000 and 1000 respectively. Assume the base for logarithm to be 10.

2+2+2+2+2+5=15 marks