IT550 - Introduction to Information Retrieval
1st Insemester Examination

29 August 2024

Total Marks: 20 **Total Time: 1 Hour**

Select exactly one option for each question below (5 Marks)

In the answer sheet clearly write the question number and the selected option for each question. You do not need to copy the question.

---

1. Consider the following examples of processed words. Identify which example(s) can be produced only as a result of lemmatization and not stemming:
   Case 1: "running" → "run" **Case 2:** "better" → "good" **Case 3:** "flies" → "fli"

   A. Case 1 only B. Case 2 only C. Case 3 only D. Both Case 1 and 2

2. Which of the following is a potential limitation of the Query Likelihood Model when dealing with queries that contain highly ambiguous terms?

   A. The model tends to favour documents containing a higher overall term frequency, regardless of the query terms.
   B. It may struggle to distinguish between different senses of an ambiguous query term, leading to a uniform likelihood across documents containing the term.
   C. The model assumes that all terms in the query are equally important, which may not be true for ambiguous terms.
   D. The Query Likelihood Model fails to account for term dependencies, treating each term independently, which can be problematic for ambiguous terms.

3. Pivoted document length normalization in the Vector Space Model is primarily designed to:

   A. Reduce the impact of document length on term weighting.
   B. Prevent the unfair penalization of longer documents.
   C. Minimize the bias toward short documents in ranking.
   D. Adjust the term frequency based on document length.

4. Query expansion is a method for adding new terms to existing queries by analyzing relevant documents. However, this is based on an inherent assumption, which can adversely affect the search performance if incorrect. What is that assumption?

   A. Substantial difference between term distributions of relevant and non-relevant documents
   B. Top-K retrieved documents are usually relevant
   C. Binary independence assumption while creating a query likelihood model
   D. Both A and B

5. Given two documents D_1 and D_2 with the following term frequencies, and using Jelinek-Mercer smoothing (also known as linear interpolation smoothing) with $\lambda = 0.7$, what is the query likelihood $P(q|D_1)$ for the query q = "apple orange"?

   - **Document** D_1: "apple" occurs 3 times, "orange" occurs 1 time, total terms = 10.
   - **Document** D_2: "apple" occurs 1 time, "orange" occurs 2 times, total terms = 10.
   - **Collection** C: "apple" occurs 10 times, "orange" occurs 5 times, total terms = 100.

   A. $0.7 \times (\frac{3}{10}) + 0.3 \times (\frac{10}{100})$

   B. $0.7 \times (\frac{3}{10}) \times (\frac{1}{10}) + 0.3 \times (\frac{10}{100}) \times (\frac{5}{100})$

   C. $(0.7 \times \frac{3}{10} + 0.3 \times \frac{1}{10}) \times (0.7 \times \frac{1}{10} + 0.3 \times \frac{5}{100})$

   D. $(0.3 \times \frac{3}{10} + 0.7 \times \frac{10}{100}) \times (0.3 \times \frac{1}{10} + 0.7 \times \frac{5}{100})$

Answer the following questions

Include only a reasonable amount of details or intermediate calculations to support your answers. Length or answers and marks obtained are independent variables.

6. Consider a Probabilistic Information Retrieval model where you are given the following information - In a collection of 10,000 documents, 500 are known to be relevant to a particular query. A term t appears in 100 of the relevant documents and in 400 of the non-relevant documents. $P(x)$ and $P(\neg x)$ are probabilities of event x occurring and not occurring respectively. Answer the questions below: **[Total: 5 Marks]**

   a. Calculate the probability that a relevant document contains a term t $(P(t|R))$ **1 Mark**

   b. Calculate the probability that a non-relevant document contains the same term t $(P(t|\neg R))$ **1 Mark**

   c. Calculate the odds of relevance $O(R|t)$ when the term t is observed in a document. **2 Marks**

   d. Given the odds of relevance, calculate the probability that the document is relevant $P(R|t)$. **1 Mark**

7. Consider a search engine that returns the following ranked list of documents for a given query:

Rank	Document ID	Relevance Grade (0-3)
1	D1	3
2	D2	2
3	D3	0
4	D4	1
5	D5	2

Answer the questions below: **[Total: 5 Marks]**

   a. Calculate the Discounted Cumulative Gain (DCG) at rank 5. **2 Marks**

   b. Calculate the Ideal Discounted Cumulative Gain (IDCG) at rank 5. **2 Marks**

   c. Compute the Normalized Discounted Cumulative Gain (NDCG) at rank 5. **1 Mark**

8. Consider the following two documents and a query:

   - **Document 1 (D1):** "Information retrieval systems are used to model queries and find information."
   - **Document 2 (D2):** "Retrieval methods include Boolean and probabilistic models."
   - **Query (Q):** "(information OR retrieval) AND model"

   Assume the standard preprocessing steps. Known stopwords = {are, to, and}.
   Answer the questions below: **[Total: 3 Marks]**

   a. Convert the two documents into Boolean vectors using appropriate vocabulary. **1 Marks**
   b. Convert the query into Boolean vector using appropriate vocabulary. **1 Marks**
   c. Determine which document(s) match the query using the Boolean model. **1 Mark**

9. In query likelihood model with linear smoothing the ranking of a document is given by the equation below:

$$P(d|q) = \prod_{t \in q} (\lambda \cdot P(t|M_d) + (1 - \lambda) \cdot P(t|M_c))$$

   Here the first term is the probability of a term appearing in the query assuming a given document model, and the second term is the same given a collection model. Assume the query has two terms w_1 and w_2, document D1 has two terms w_1 and w_3 and document D2 has w_1 and w_4. Each document has a single term that overlaps with the query. However, the scores for each document will turn out to be different based on the $P(t|M_c)$ for w_3 and w_4. Is this ranking fair? Provide reasoning for the same in no more than 5-7 sentences. **[2 Marks]**