Task 2.1 – Classical Text Retrieval (multiple choices)

Select the correct answer for the following questions.

1) During the online querying phase, which of the following is the primary task in assessing a document's relevance to a query?
   A. Correcting spelling mistakes in the query.
   B. Comparing features of the document to the query.
   C. Indexing new documents for future use.
   D. Training of improved token generators.

2) What is the main challenge during the offline indexing phase?
   A. Correcting spelling mistakes in the documents.
   B. Finding the documents to index (web crawler).
   C. Extracting meaningful features from text documents.
   D. Ranking documents based on retrieval status value (RSV).

3) What is the purpose of dividing the search into an offline indexing phase and an online querying phase?
   A. To train neural networks for fast semantic search.
   B. To store documents for future use.
   C. To speed up searches for queries.
   D. To comply with data privacy regulations.

4) What is the primary reason for splitting documents into smaller parts in information retrieval systems, as mentioned in the course material?
   A. To improve the accuracy of search results for complex queries.
   B. To save storage space for large documents.
   C. To speed up the indexing of documents.
   D. To fit text parts into large language models.

5) What is the primary purpose of stemming in text tokenization and linguistic transformation?
   A. To create compound words.
   B. To reduce tokens to a common stem for better term matching.
   C. To reduce the size of the vocabulary to accelerate retrieval.
   D. To add special characters to words.

Task 2.1 – Classical Text Retrieval (multiple choices)

Select the correct answer for the following questions.

6) What is the purpose of the inverse document frequency (IDF) in text summarization?
   A. To assign each term a dimension in the feature space.
   B. To preserve term frequencies in the feature vector.
   C. To automate stop word removal.
   D. To differentiate the significance of terms.

7) What is the key difference between the set-of-words model and the bag-of-words model in feature vector creation?
   A. The set-of-words model considers term frequencies, while the bag-of-words model does not.
   B. The set-of-words model preserves the order of terms, while the bag-of-words model does not.
   C. The set-of-words model uses binary feature vectors, while the bag-of-words model preserves term frequencies.
   D. The bag-of-words model is for offline analysis, the set-of-words model is fo online analysis.

8) Most text retrieval models including state-of-the-art ones like BM25 treat terms as independent from each other to improve retrieval speed. What is the biggest challenge with the independence assumption?
   A. It leads to the over-stemming of terms impacting the precision of queries.
   B. It makes documents more similar to each other impacting the recall of queries.
   C. It makes it more difficult to find semantically close documents and to improve recall.
   D. It requires longer queries to find relevant documents and to improve precision.

9) Why is the common practice today to retain all terms, including stop words, and how can we prevent their negative impact on search results?
   A. To increase storage space efficiency accelerating retrieval speed.
   B. To increase vocabulary size so that inverted files become more effective and we find more relevant documents.
   C. To maintain the ability to search for all possible queries, but reducing their impact on relevance scores.
   D. Stop words should always be removed. They only waste resources.

Task 2.1 – Classical Text Retrieval (multiple choices)

Select the correct answer for the following questions.

10) What is one of the disadvantages of the Boolean model of information retrieval?
    A. Users are looking for more advanced search experiences.
    B. It offers precise control for including or excluding documents.
    C. It can explain why a document was considered relevant.
    D. Users may find it hard to express a complex information need as a combination of Boolean operators.

11) What is the primary difference between the inner vector product and the cosine measure in the context of vector space retrieval?
    A. Inverted files only work with vector product while we need full scans for the cosine measure.
    B. The inner vector product produces values between 0 and 1, while the cosine measure can produce values outside this range.
    C. The inner vector product considers both term frequencies and inverse document frequency (idf), while the cosine measure only considers term frequencies.
    D. The inner vector product focuses on term occurrences in documents, while the cosine measure focuses on directions.

12) Why is the vector space retrieval model able to provide partial match results, and how does it help users
    A. IDF splits the document collection into several cluster with partial overlap to the query; this accelerates retrieval times.
    B. Documents only need to match with at least one query term to have a positive score; users do not have to adjust queries to get results.
    C. By eliminating stop words, the model can provide partial semantic matching; user can focus on terms with high IDF values
    D. The confusion matrix is only partially matched, allow users to better optimize for precision.

13) How is the issue of similar query terms, like 'house' and 'villa,' typically addressed in vector space retrieval?
    A. By automatically expanding queries with related terms.
    B. By not allowing similar query terms to be used in the same query.
    C. By disregarding similar query terms altogether.
    D. By gathering feedback from users (“did you mean?”)

Task 2.1 – Classical Text Retrieval (multiple choices)

Select the correct answer for the following questions.

14) What is the primary advantage of using inverted indexes in information retrieval systems when compared to a simple storage approach that scales linearly with collection and vocabulary size?

    A. Inverted indexes consume less storage space.
    B. Inverted indexes allow for more efficient compression of term frequencies.
    C. Inverted indexes enable faster processing by reading only relevant data.
    D. Inverted indexes maintain a more orderly structure for document storage.

15) Why is evaluating queries like "cat OR NOT(dog)" not considered ideal in the described query evaluation method using sorted postings?

    A. The method requires listing all documents except those in the 'dog' postings, which can be slow for terms with low document frequencies.
    B. Evaluating such queries is computationally infeasible due to the complexity of the Boolean expression.
    C. The method lacks support for nested NOT operators, making it challenging to handle queries with multiple operands.
    D. Queries with OR and NOT operators are conceptually confusing for users (so-called or-not confusion) and thus not used in retrieval.

16) In the Binary Independence Retrieval (BIR) model, what is the primary difference between the document-at-a-time (DAAT) and term-at-a-time (TAAT) retrieval methods?

    A. DAAT retrieves all documents matching one query term at a time, while TAAT retrieves all query terms for each document.
    B. DAAT focuses on processing documents with the most recent updates, while TAAT processes documents with the least recent updates.
    C. DAAT computes scores for documents in a single step and maintains a top-k list, while TAAT requires a sizable scores dictionary.
    D. DAAT processes several queries concurrently, TAAT processes one query but several documents at a time.

Task 2.1 – Classical Text Retrieval (multiple choices)

Select the correct answer for the following questions.

17) In Lucene, what is the primary purpose of marking documents as deleted instead of physically deleting them in the indexing process?
    A. To save storage space in the index.
    B. To ensure that documents are removed from the index immediately.
    C. To ensure data integrity while concurrently updating and searching.
    D. To simplify the retrieval process by keeping only the most recent version.

18) What is the purpose of using sharding in distributed search systems like Solr, Elasticsearch, and OpenSearch?
    A. To increase query speed for individual documents.
    B. To ensure exact consistency in scores for identical documents.
    C. To distribute collections across multiple worker nodes.
    D. To eliminate the need for segment-level map-reduce

19) What is one of the key benefits of using shard replication in distributed search systems like Solr, Elasticsearch, and OpenSearch?
    A. Reducing the number of shards in the system.
    B. Enhancing the speed of indexing new documents.
    C. Increasing overall system availability and resilience against failures.
    D. Eliminating the need for leader nodes in the search coordination

20) What is one of the primary advantages of deploying search clusters across multiple regions for a search application?
    A. Reducing the number of documents in the search index.
    B. Decreasing the need for shard replication.
    C. Improving the latency for query results.
    D. Minimizing the need for DNS routing

Task 2.2: Vector Space Retrieval (theoretical)

In the script, we employ inner vector products and cosine measures to rank documents based on their similarity to the query. Our current task involves examining the "semantics" of these functions from a geometric standpoint. We start by looking at a query with just two terms and then extend the analysis to higher dimensions for simplicity.

a) Begin by examining a two-term query and defining a similarity threshold α. For both metrics, pinpoint the subset of documents with a similarity score surpassing α. Characterize this subset in geometric language.

b) Using the geometric semantics from part a), determine the documents favored by the measures. Create an example document that "wins" the search with the highest scores. Extend this analysis to queries with more than two terms.

c) In web search, queries are frequently brief. What occurs when you choose just one query term? Do the measures function effectively in this extreme scenario?

We aim to conduct text similarity searches, like finding pages with plagiarized content. We employ the bag-of-words model and measure the texts' similarity using Euclidean distance. Let $\boldsymbol{q}$ represent the term vector for Query Q, and $\boldsymbol{d}$ be the term vector for document D. Then:

$$\delta(Q,D) = \sqrt{\sum_i (q_i - d_i)^2}$$

In contrast to the inner vector product and the cosine measure, small distances are better (more relevant) than large distances (less relevant).

d) Similar to a), explain the subspace of documents having a distance of at most β from query Q. Identify the documents that rank highest using this distance measure. Does this approach work for our scenario of finding similar pages, and if so, why?

Task 2.3: Probabilistic Retrieval (theoretical)

In this task, we examine the binary independence retrieval (BIR) model and illustrate the approach with straightforward examples.

a) For a query Q, the BIR method provides this initial list of documents:

rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
x_2	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0
relevance	R	R	R	R	N	R	R	R	R	N	N	R	R	R	N	N	N	R	N	N

In the table, rows x_1 and x_2 show the binary representation of the 20 retrieved documents. The final row indicates user relevance assessments for each document (R for relevant, N for non-relevant). Calculate updated c_j values based on the feedback and determine the new ranking.

b) The BIR model relies on three assumptions. We will now assess their validity. To do so, we calculate the probability $P(R|\boldsymbol{x})$ using the example data from a) in two methods: 1) Count the occurrences of relevance/non-relevance for documents with representation $\boldsymbol{x}$ and compute the probability. And 2) Formulate a $P(R|\boldsymbol{x})$ equation based on r_j and n_j, as outlined in the script. We begin with the following statement:

$$sim(Q, D_i) = \frac{P(R|D_i)}{P(NR|D_i)} = \frac{P(R|D_i)}{1 - P(R|D_i)} = \frac{P(R|\boldsymbol{x})}{1 - P(R|\boldsymbol{x})} = \cdots$$

and solve for $P(R|\boldsymbol{x})$. What do you observe? Which assumption fails?

c) Examine the documents (c1-c5, m1-m4) and the query "human computer interaction." Perform two iterations with the BIR model (initialization step and one feedback step). Assume c1-c5 are relevant, while m1-m4 are non-relevant. Does the feedback step enhance retrieval? What measures can be taken to notably improve retrieval performance with feedback?

c1 **Human** machine interface for Lab ABC **computer** applications
c2 A survey of user opinion of **computer** system response time
c3 The EPS user interface management system
c4 System and **human** system engineering testing of EPS
c5 Relation of user-perceived response time to error measurement

m1 The generation of random, binary, unordered trees
m2 The intersection graph of paths in trees
m3 Graph minors IV: Widths of trees and well-quasi-ordering
m4 Graph minors: A survey

Task 2.4: Searching with Lucene (practical)

In this exercise, we use Lucene and its fuzzy retrieval model to search for movie titles. You can find the data set with 1000 movie titles on the course page and on Kaggle.com: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows.

Prioritize functionality and maintain a simple text-based interface. Offer a menu within the program or use arguments for your program to invoke the different sub-tasks. This task is open-ended, and the following sub-tasks are suggestions for what you can achieve with Lucene. Use timeboxing (allocate a specific number of hours) and stop once you encounter difficulties. Collaborate with fellow students to share strategies for tackling the challenges, as navigating through the Lucene documentation can be time consuming.

a) **Import the CSV data into a Lucene index.** Create a Lucene index and implement code to read the CSV file, extracting relevant information (e.g., movie titles, descriptions, ratings) and adding them to the index.

b) **Implement a basic search function to retrieve movies based on keywords.** Create a simple search function that allows users to input keywords and retrieve movies that contain those keywords in their titles or descriptions.

c) **Enhance the search results by considering more relevance factors.** Consider factors like movie ratings, release dates, and occurrences of keywords in different fields to improve the ranking of search results.

d) **Enhance query term matching with query expansion.** Enable fuzzy keyword matches and expand queries if not sufficient search results are provided (or trigger by interface to expand query automatically).

e) **Implement faceted search to allow users to filter results.** Create facets for genres, release years, or other relevant categories to enable users to narrow down search results. Use a feedback prompt to ask users to further filter results

f) **Add spell-checking**. Implement features that correct misspelled queries.

g) **Implement pagination for search results.** Divide search results into pages, allowing users to navigate through multiple pages of results. You can output 10 results and prompt for ‘next’ or ‘prev’ to navigate.