MCQs

1. What is an inverted index primarily used for in Information Retrieval?
   a) Data encryption
   b) Query optimization
   c) Web design
   d) User authentication

2. What represents a term and a document in a term-document incidence matrix?
   a) Row - Term, Column - Document
   b) Row - Document, Column - Term
   c) Row - Query, Column - Term
   d) Row - Term, Column - Query

3. In Blocked Sort-Based Indexing, what type of data structure is commonly used for sorting smaller blocks of data?
   a) B-tree
   b) Binary heap
   c) Hash Table
   d) External Sorting

4. What challenge does the sparsity of term-document incidence matrices present?
   a) Decreased computational complexity
   b) Difficulty in data visualization
   c) Higher storage requirements
   d) Easier data manipulation

5. Which retrieval model involves using logical operators for query processing?
   a) Vector Space Model
   b) Probabilistic Model
   c) Boolean Retrieval Model
   d) Neural Network Model

6. In the context of information retrieval, 'stemming' is used to:
   a) Increase the number of relevant documents retrieved.
   b) Reduce a word to its base or root form.
   c) Encrypt sensitive data.
   d) Visualize data patterns.

7. What is the primary purpose of using an n-gram model in text indexing?
   a) To detect grammatical errors in text.
   b) To handle phrase queries and approximate string matching.
   c) For encrypting text data.
   d) To increase the speed of text rendering.

8. If we are using a linked list for storing the inverted indices over a corpus of |D| documents, then the length of the smallest linked-list will be X, and length of the longest linked-list could be Y.
    a) X = 0 , Y = |D|
    b) X = 1, Y = |D|-1
    c) X = 1, Y=|D|
    d) X = 1, Y = |D|-1

MSQs

1. Which of the following statements are true about BSBI and SPIMI in the context of information retrieval?

a) BSBI is efficient for small datasets as it requires all data to fit into memory.
b) SPIMI is designed to handle large datasets by processing data in a single pass and utilizing in-memory operations.
c) In BSBI, data is sorted and indexed in blocks which are then merged, making it suitable for large datasets that do not fit into main memory.
d) SPIMI relies heavily on external sorting, similar to BSBI, to manage data sorting and indexing.

2. Consider a sentence of length m (having m tokens). The total possible number of n-grams possible from the given sentence:

a) m^2
b) C^{m}_{2} + m
c) m*(m+1)/2
d) m*(m-1)/2

Fill in the blanks

1. Information Retrieval is a _____ model whereas data retrieval is a ____ model
2. The ___ boolean operator is used for intersection and ___ boolean operator is used for union
3. If the list lengths are x and y, the merge takes O___ operations.
4. ___ still uses a boolean queries for retrieval

Answers:

1) Probabilistic, deterministic
2) AND, OR
3) (x+y)
4) Email/library catalog/macOS Spotlight/Westlaw

<u>Subjective</u>

Q1. Given Inverted Index data structure with any 6 terms in 25 documents below

a: 1, 2, 3, 5, 8, 10
b: 1, 4, 6, 8, 9, 11, 13, 15
c: 2, 4, 7, 12, 14, 16, 17, 19
d: 3, 5, 6, 8, 11, 14, 17, 20, 22
e: 7, 9, 10, 12, 15, 18, 21, 23
f: 10, 13, 16, 18, 20, 24, 25

Give the boolean expression as well as the output for

a) Which documents contain both terms "a" as well as "b"?
b) Which documents contain the term "e" but not the term "f"?
c) Which documents contain either the term "c" or "d", but not both?

Ans.

a) a AND b → 1,8
b) e AND NOT f → 7, 9, 12, 15, 21, 23
c) (c or d) AND NOT (c AND d) → 2,3,4,5,6,7,8,11,12,16,19,20,22

Q2. What is dynamic indexing? Why is it needed? Describe its approach.

Ans. Dynamic indexing is a way of indexing wherein the information retrieval systems continuously update and expand the index as new documents are added to a collection or existing documents are modified or deleted.
It is needed because collections are not static and Documents come in over time and need to be inserted, deleted and modified.
<u>Approach</u>

- Maintain “big” main index
- New docs go into “small” auxiliary index
- Search across both, merge results
- Periodically, re-index into one main index