1)
Ms Alice went to the Sarojini Market with her assistant Mr Bob. While the lady was bargaining, Bob noted down every utterance between the shopkeeper and Alice in his mobile phone. For example: “*Alice: 500 rupees is very costly bhaiya*”. “*Shopkeeper: no madam it is fine.*” “*Alice: no bhaiya …*”. This is an example of **X**. Later Bob came home and annotated each dialogue with the shop location. For example: “shop#1[*Alice: 500 rupees is very costly bhaiya…*], shop#2[*Alice: how much for …*], …”. This is an example of **Y**. Then for each product Bob recorded the asking price of each shop in a spreadsheet. This is an example of **Z**. Then which of the following are correct?

a) X is unstructured data, Y is semi-structured data, and Z is structured data.
b) X and Y are semi-structured data, whereas Z is structured data. Explanation.
c) If Bob adds proper punctuation marks to X then it will become semi-structured data from unstructured data. Explanation.
d) If Bob removes character attributions (*Alice:* , *Shopkeeper:*) from X then it will become/remain unstructured data. Explanation.

2)
Given that the query (q) and database (db) are passed through the same (deterministic) representation system. If we use exact matching (0/1) as the similarity measure between the query and each item of the database, then which of the following holds true?

a) It is a typical example of Information Retrieval. Explanation.
b) Ranking of the results will be dependent on the sequence of matching. E.
c) After the user relevance feedback, the ranked list of items will not change. E.
d) Number of retrieved items will not change for a given query over time. E.

3)
Which of the following properties holds true for information retrieval using Term-document Indices Matrix (M)?

a) For a large vocabulary, the memory requirements of such a matrix is very high.
b) The matrices are generally very dense. Explanation.
c) Given no memory constraints, the retrieval happens in O(1) time.
d) By taking word representations from M, the cosine similarity between two words is in the range of -1 to 1. Explanation.

4)
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

5)
Consider the following paragraph: *Shyama’s father had great courage. He had three wives.*
Which of the following statements are true?

a) Number of tokens are between 10 to 12 (including both). Explanation.
b) Number of stopwords are between 8 to 10 (including both). Explanation.

c) There exists at least two words whose lemma is not the same as its stem. Explanation.

d) The size of vocabulary in this paragraph equals the number of tokens in this paragraph. Explanation.

Descriptive type question

What are the advantages of using stemming over lemmatization and vice versa? Explanation

Ans:

Stemming > Lemmatization

i) Stemmers can be trained without any linguistic knowledge hence more **robust** i.e. we **don't need any linguistic expert** to train a stemmer.

ii) Stemming is slightly **faster** than lemmatization.

Lemmatization > Stemming

i) The root word produced by lemmatizers **exists in the language vocabulary** i.e. their output (lemma) is comprehensible.

ii) Lemmatization is **context-dependent**.

iii) Lemmatization is **more accurate** than stemming i.e. number of times a lemmatizer will make an error will be smaller than the number of times a stemmer will make an error.

iv) Stemmer **will not work correctly for natural languages that** are not space separated (like old Chinese) or that are synthetic in nature (like Arabic) where the word spelling changes with an inflection (for example see → saw, good → better).

Source1

Arunim

6) Which of the following search systems still use Boolean queries for retrieval:

a) Mac OS spotlight

b) Cortona Search

c) Westlaw

d) Google scholar

Reason: slides

7) 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

Reasons both are same mathematically

8) Shashank wants to make a search engine using phase queries. During his process, he was

unable to get the desired performance and results using the extended n-words indexing. Which of the following reasons could be the cause of it

a) Extended n-words indexing results in True Negatives which can give incorrect results
b) Extended n-words indexing results in False Positives which can give incorrect results
c) Dictionary blows up for n greater than 2
d) All of the above

Reasons Trust Negatives are not desired and they are incorrect results, for n_word greater than 2, it will increase the computation to large extent

9) Given a positional index for "**be"** containing documents from 1 to 5 :

be: 993427;
1: 7, 18, 33, 72, 86, 231;
2: 3, 149;
3: 17, 191, 291, 430, 434;
4: 363, 367, 504, 893, 997;
5: 102, 103, 104, 441, 562;

Select the documents containing the proximity phrase: "***be \4 be***"

a) Document 1
b) Document 3
c) Document 4
d) Document 5

Reasons All the documents, within the different of 4 positional index will considered, which means, between 2 be's, at max 4 words can come

10) Time complexity of merging 2 linked list of size x and y:

a) O(x+y)
b) O(min(x,y))
c) O(max(x,y))
d) Θ(min(x,y))

Reasons For the given question, mathematically, O(x+y) and O(max(x,y)) will correspond to same time complexity

<u>Numerical Question</u>
Given Inverted Index data structure with any 6 terms in 25 documents below

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

f - 1, 3, 4, 7, 8, 10, 14, 16, 22, 25

a) Let suppose every integer occupies 4 bytes and the rest of the space covered by the data structure is S. Find the total space occupied by the **complete** inverted index data structure in terms of S.

   Ans
   ans: (S + (50+6)*4)
   6 for size of each list
   50 after adding the space occupied by each document id.

   if it's S+50*4, you can give them half a mark as long as they have written the explanation and assumption.

   if the answer is with an assumption, then please correct the copies according to what you feel is the correct answer.

b) Compute:
   (a OR b) AND (e OR f)
   Ans
   4, 7, 10, 12, 14, 22

c) Compute:
   c AND NOT d

   Ans
   null set