(2 ½ Hours) **[Total Marks: 75]**

N.B. 1) All questions are compulsory.
2) Figures to the right indicate marks.
3) Illustrations, in-depth answers and diagrams will be appreciated.
4) Mixing of sub-questions is not allowed.

Q. 1 Attempt All

(a) Select the correct alternative from the options given: **(10M)**

(i) Information retrieval is querying of _____ textual data.
(a) Structured (b) Unstructured
(c) Formatted (d) Unformatted

(ii) _____ is what fraction of the relevant documents in the collection were returned by the system.
(a) Index (b) Inverted index
(c) Precision (d) Recall

(iii) In a _____, the dictionary contains all k-grams that occur in any term in the vocabulary.
(a) Indexing (b) Permuterm index
(c) Permuterm (d) k-gram index

(iv) Which of the following is a technique for context sensitive spelling correction?
(a) The Jaccard Coefficient (b) Soundex algorithms
(c) Levenshtein distance (d) k-gram indexes

(v) _____ takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
(a) Map (b) Reduce
(c) Map Reduce (d) Collections

(vi) _____ in Information retrieval are short fragments of text extracted from the document content or its metadata.
(a) snippets (b) question answering
(c) long document (d) metadata of document

(vii) _____ problem means that there is a need to be enough other users already in the system to find a match
(a) unique taste (b) active users
(c) sparity (d) Cold start

(viii) _____ approaches are commonly used for data collections with complex structures that mainly contain nontext data.
(a) document-centric (b) data-centric
(c) text-centric (d) query-centric

(ix) The standard for accessing and processing XML documents is the XML _____

(a) Document Oriented Model (b) Database Object Model
(c) Data Object Model (d) Document Object Model

(x) _____ refers to a huge database of internet resources such as web pages, newsgroups, programs, images etc.

(a) result page (b) search engine
(c) database (d) web crawler

(b) **Fill in the blanks by selecting from the pool of options:** **(5M)**

(Vector Space Model, Xtensible Markup Language, Extensible Markup Language, collaborative filtering, Hyperlink, SEO, Levenshtein distance, hub, authority, surface web, query suggestion)

(i) When we replace a character of a string by another character, it is called as _____

(ii) A good _____ page for a topic links to many authority pages for that topic

(iii) _____ forms a directed edge from one node to another node in a web graph

(iv) XML stands for _____

(v) _____ is the algebraic model for representing text documents as vectors of identifiers

Q. 2 Attempt the following (Any THREE) **(15M)**

(a) Explain inverted index used in IR with the help of an example.

(b) Explain the following terms
- i) Corpus
- ii) Precision
- iii) Recall
- iv) Stop words
- v) Token

(c) Draw the term document incidence matrix for the following document collection and answer the given queries

Doc 1 new home sales top forecasts
Doc 2 home sales rise in july
Doc 3 increase in home sales in july
Doc 4 july new home sales rise

- i) Home and sales and july
- ii) Rise and sales not increase

(d) Explain the search process using Binary tree.

(e) Explain the SOUNDEX algorithm used in Phonetic correction

(f) What is spelling correction? State and explain its different forms.

Q. 3 Attempt the following (Any THREE) **(15M)**

(a) Explain hubs and authorities in detail.

(b) How page rank algorithm is used for ranking webpages?

(c) Explain how term frequency and inverse document frequency can be used in ranking web pages?

(d) Explain the HDFS architecture.
(e) State the advantages and disadvantages of personalized search.
(f) What is snippet? Explain its importance in information retrieval.

Q. 4 Attempt the following (Any THREE) **(15)**

(a) Explain the following terms in SPAM.
   i) Content hiding
   ii) Cloaking
   iii) Redirection
   iv) URL spamming
   v) Term spamming

(b) What is user query? State and explain the different types of queries entered by the user.
(c) What is Web search architecture? Explain its components.
(d) How Vector space model is used for information retrieval?
(e) Explain the Data Centric XML retrieval with the help of examples.
(f) What is XML retrieval system? State and explain the challenges of XML.

Q. 5 Attempt the following (Any FIVE) **(15)**

(a) State the challenges in information retrieval.
(b) Compute the Edit distance to convert CATS to FATS.
(c) Explain invisible web.
(d) What is the need of question answering system?
(e) What is Collaborative filtering?
(f) Explain the indexing process in search engine.
(g) Explain the following:
   i) Web graph
   ii) Static web pages
   iii) Web size
(h) What is black hat SEO?