Q. 1 Attempt All (Each of 5Marks) (15M)

(a) Multiple Choice Questions (5M)

i) ______ are indexed units in incidence matrix.
   a. Terms
   b. Collection
   c. Information
   d. Data

ii) The number of documents in the collection that contain a term t is called as _______
    a. Document Index di_t
    b. Document frequency df_t
    c. Document Inverse din_t
    d. Document Incidence Matrix dim_t

iii) The standard way of quantifying the similarity between two documents d1 and d2 is to compute the _______ of their vector representations.
     a. sine similarity
     b. cot similarity
     c. cosine similarity
     d. None

iv) CPM stands for __________________
    a. Cost per mil
    b. Cost per making
    c. Cost per manage
    d. Cost per migrating

v) ______ fraction of the returned results are relevant to the information need.
   a. Proximity
   b. Posting Merge
   c. Posting list
   d. Precision

(b) Fill in the blanks (5M)

(in-links, Static, semistructured, Document Object Model, two)

i) IR is also used to facilitate ________ search such as finding a document where the title contains Java and the body contains threading.

ii) ________ web pages are those whose content does not vary from one request for that page to the next.

iii) Every web page is assigned ________ scores.

iv) The standard for accessing and processing XML documents is the XML ________.

v) The hyperlinks into a page as ________.
Q. 2 Attempt the following (Any THREE)(Each of 5Marks) (15M)
(a) Brief overview of Information retrieval.
(b) What are the components of Information retrieval? Explain with diagram.
(c) Brief the history of Information retrieval.
(d) List the forms of spelling correction in Information retrieval. Explain.
(e) Explain the architecture of open source engine framework.
(f) Draw the inverted index that would be built for the following document collection.
   Doc 1 one fish, two fish
   Doc 2 red fish, blue fish
   Doc 3 one red bird

Q. 3 Attempt the following (Any THREE) (Each of 5Marks) (15M)
(a) Discuss Hubs and Authorities.
(b) Explain the concept of cosine similarity with example.
(c) What is Personalized search? State factors affecting it.
(d) Explain the concept of Collaborative filtering.
(e) What is Question answering? Explain.
(f) Give the meaning of cross lingual retrieval. Analyse its process.

Q. 4 Attempt the following (Any THREE) (Each of 5Marks) (15M)
(a) Explain the terms: Web, Web pages, Web graph with example.
(b) Discuss categories of user needs in web queries for query analysis.
(c) What are the basic building blocks of Search Engine Architecture? Explain.
(d) Give the challenges in XML retrieval.
(e) Write a note on Web Size Measurement.
(f) Write a note on sponsored search.

Q. 5 Attempt the following (Any THREE) (Each of 5Marks) (15M)
(a) Compute the Levenshtein edit distance between “GUMBO” and “GAMBOL”.
(b) Give the concept of wild card queries in IR.
(c) Define Page rank. How to compute page rank for a webpage? Give example.
(d) What is MapReduce? Explain its paradigm.
(e) Differentiate between Text Centric v/s Data Centric XML.