CS8080- INFORMATION RETRIEVAL TECHIQUES

Question Bank

CS8080 INFORMATION RETRIEVAL L T P C

3 0 0 3

OBJECTIVES:

- To understand the basics of Information Retrieval.
- To understand machine learning techniques for text classification and clustering.
- To understand various search engine system operations.
- To learn different techniques of recommender system.

UNIT I INTRODUCTION 9

Information Retrieval – Early Developments – The IR Problem – The Users Task – Information versus Data Retrieval – The IR System – The Software Architecture of the IR System – The Retrieval and Ranking Processes – The Web – The e-Publishing Era – How the web changed Search – Practical Issues on the Web – How People Search – Search Interfaces Today – Visualization in Search Interfaces.

UNIT II MODELING AND RETRIEVAL EVALUATION 9

Basic IR Models – Boolean Model – TF-IDF (Term Frequency/Inverse Document Frequency) Weighting – Vector Model – Probabilistic Model – Latent Semantic Indexing Model – Neural Network Model – Retrieval Evaluation – Retrieval Metrics – Precision and Recall – Reference Collection – User-based Evaluation – Relevance Feedback and Query Expansion – Explicit Relevance Feedback.

UNIT III TEXT CLASSIFICATION AND CLUSTERINGNG 9

A Characterization of Text Classification – Unsupervised Algorithms: Clustering – Naïve Text Classification – Supervised Algorithms – Decision Tree – KNN Classifier – SVM Classifier – Feature Selection or Dimensionality Reduction – Evaluation metrics – Accuracy and Error – Organizing the classes – Indexing and Searching – Inverted Indexes – Sequential Searching – Multi-dimensional Indexing.

UNIT IV WEB RETRIEVAL AND WEB CRAWLING 9

The Web – Search Engine Architectures – Cluster based Architecture – Distributed Architectures – Search Engine Ranking – Link based Ranking – Simple Ranking Functions – Learning to Rank – Evaluations — Search Engine Ranking – Search Engine User Interaction – Browsing – Applications of a Web Crawler – Taxonomy – Architecture and Implementation – Scheduling Algorithms – Evaluation.

UNIT V RECOMMENDER SYSTEM 9

Recommender Systems Functions – Data and Knowledge Sources – Recommendation Techniques – Basics of Content-based Recommender Systems – High Level Architecture – Advantages and Drawbacks of Content-based Filtering – Collaborative Filtering – Matrix factorization models – Neighborhood models.

TOTAL: 45 PERIODS

OUTCOMES:

Upon completion of the course, students will be able to:

- Use an open source search engine framework and explore its capabilities
- Apply appropriate method of classification or clustering.
- Design and implement innovative features in a search engine.
- Design and implement a recommender system.

TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, ―Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.
2. Ricci, F, Rokach, L. Shapira, B.Kantor, ―Recommender Systems Handbook, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, ―Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, ―Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

CO1	Remember the basics of Information Retrieval
CO2	Analysis Modeling And Retrieval Evaluation in Information retrieval
CO3	Understand machine learning techniques for text classification and clustering and Create various search engine system operations.
CO4	Analysis Web Retrieval And Web Crawling in Information Retrieval
CO5	Learn different techniques of recommender system
CO6	Evaluate the filtering and collaborative filtering models

UNIT I INTRODUCTION

Information Retrieval – Early Developments – The IR Problem – The User's Task – Information versus Data Retrieval - The IR System – The Software Architecture of the IR System – The Retrieval and Ranking Processes - The Web – The e-Publishing Era – How the web changed Search – Practical Issues on the Web – How People Search – Search Interfaces Today – Visualization in Search Interfaces.

PART - A

1. **Give any two advantages of using artificial intelligence in information retrieval tasks. (Apr/May 2018) U**

   The advantages of using artificial intelligence in information retrieval tasks are as follows:

   - Information characterization
   - Search formulation in information seeking
   - System Integration
   - Support functions

2. **How can IR be studied from complementary points of view? U**

   IR can be studied from two rather distinct and complementary points of view:

   - Computer-centered
     - Building up efficient indexes
     - Processing the user queries with high performance
     - Developing ranking algorithms to improve the results
   - Human-centered
     - Studying the behavior of the user
     - Understanding their main needs
     - Determining how such understanding affects the organization and operation of the retrieval system

3. **How does the large amount of information available in web affect information retrieval system implementation? (Apr/May 2018) U**

   Large amount of unstructured designed information is difficult to deal with. Obtaining specific information is a hard mission and takes a lot of time. Information Retrieval System (IR) is a way to solve this kind of problem. IR is a good mechanism but does not give the perfect solution. It can cause the system to

   - Information overload
   - Time consuming

4. **Define Information Retrieval. (Nov/Dec 2016, Apr/May 2022) R**

   Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories,

particularly textual information. Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.

5. **Enumerate the purpose of Information Retrieval. (Apr/May 2021) AN**
   The major objective of an information retrieval system is to retrieve the information – either the actual information or the documents containing the information – that fully or partially match the user's query.

6. **What are the functions of information retrieval system? (Nov/Dec 2020) U**
   An information retrieval system is designed to analyze, process and store sources of information and retrieve those that match a particular user's requirements.

7. **List the issues in information retrieval system. U**
   The issues of IR systems are:
   - Assisting the user in clarifying and analyzing the problem and determining information needs.
   - Knowing how people use and process information.
   - Assembling a package of information that enables group the user to come closer to a solution of his problem.
   - Knowledge representation.
   - Procedures for processing knowledge/information.
   - The human-computer interface.
   - Designing integrated workbench systems
   -

8. **Specify the role of an IR system. (Nov/Dec 2016) U**
   Information retrieval is fast becoming the dominant form of information access which covers various kinds of data and information problems. Automated information retrieval systems are used to reduce what has been called "information overload". The public libraries use IR systems to provide access to books, journals and other documents.

9. **Differentiate information and data retrieval. (Apr/May 2021) AN**

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user

	query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.
Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics.
Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.
The results obtained are approximate matches.	The results obtained are exact matches.
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

10. List few Information Retrieval Models. (Nov/Dec 2020) R

Boolean, Vector and Probabilistic are the three classical IR models.

11. What is open source search framework? List examples. (Nov/Dec 2020) R

A search engine is a software program that helps people find the information they are looking for online using search queries containing keywords or phrases. There exist some popular open-source search engines which can be used to build search functionality in your website.

- Apache Lucene
- Apache Solr
- Elasticsearch
- MeiliSearch
- Typesense

12. Identify the types of search engines. (Apr/May 2021) U

Mainstream search engines like Google might be top of mind when we think about search engines, but there are other types of search engines that allow us to navigate the internet.

- Mainstream search engines. Mainstream search engines like Google, Bing, and Yahoo! are all free to use and supported by online advertising. They all use variations of the same strategy (crawling, indexing, and ranking) to let you search the entirety of the internet.
- Private search engines. Private search engines have risen in popularity recently due to privacy concerns raised by the data collection practices of mainstream search engines. These include anonymous, ad-supported search engines like DuckDuckGo and private, ad-free search engines like Neeva.
- Vertical search engines. Vertical search, or specialized search, is a way of narrowing your search to one topic category, rather than the entirety of the web. Examples of vertical search engines include:
  1. The search bar on shopping sites like eBay and Amazon
  2. Google Scholar, which indexes scholarly literature across publications
  3. Searchable social media sites and apps like Pinterest

- Computational search engines. WolframAlpha is an example of a computational search engine, devoted to answering questions related to math and science.

13. What are the major impacts that Web has had in the development of IR? R

The major impacts are:

- Web Document Collection and Search Engine Optimization
- Size of the collection and the volume of user queries submitted on a daily basis
- Predicting relevance is much harder than before due to the vast size of the document collection.
- Web is not just a repository of documents and data, but also a medium to do business.
- Web advertising and other economic incentives.

14. What are the performance measures of search engine? (Nov/Dec 2021) R

The two fundamental metrics are recall, measuring the ability of a search engine to find the relevant material in the index, and precision, measuring its ability to place that relevant material high in the ranking.

15. List the challenges of searching for information on the web. U

The challenges are listed as below.

- Size of the databases, Web coverage
- Up-to-dateness of search engines' databases:
- Web content
- The Invisible Web
- Spam

16. How Search Engine Works? Explain. AN

Search engines have two major functions: crawling and building an index, and providing search users with a ranked list of the websites.

1. Crawling and indexing
2. Providing answers

17. Define Indexing. R

Indexing is an important process in Information Retrieval (IR) systems. Indexing reduces the documents to the informative terms contained in them. It provides a mapping from the terms to the respective documents containing them.

18. State the role of AI in IR. AN

The Artificial Intelligence models and techniques are used in the design of a small Information Retrieval system. In particular, some knowledge representation models, such as semantic networks and frame-like structures, are viewed as interesting tools for the implementation of a thesaurus, and also for a description of the stored documents' contents.

19. Define Zipf's law. R

An empirical rule describes the frequency of the text words. It state that the ith most frequent word appears as many times as the most frequent one divided by i@, for some @>1.

20. What are the major activities of the information seeking process model? R

The classic notion of the information seeking process model as described by Sutcliffe and Ennis is formulated as a cycle consisting of four main activities:

- Problem identification,
- Articulation of information need(s),
- Query formulation, and
- Results evaluation.

21. List the advantage of open source. U

The advantages of open source are,

- The right to use the software in any way.
- There is usually no license cost and free of cost.
- The source code is open and can be modified freely.
- Open standards.
- It provides higher flexibility.

22. List the disadvantage of open source. U

The disadvantages of open source are,

- There is no guarantee that development will happen.
- It is sometimes difficult to know that a project exist, and its current status.
- No secured follow-up development strategy.

23. What are the main interactive information visualization techniques? R

The main interactive information visualization techniques include:

- Panning and zooming,
- Distortion-based views (including focus plus context), and
- The use of animation to retain context and help make occluded information visible.

Part B & C

1. Differentiate between Information Retrieval and Web Search. **(Nov/Dec 2017) AN**
2. Explain the issues in the process of Information Retrieval. **(Nov/Dec 2017) U**
3. Explain in detail, the components of Information Retrieval and Search engine.

   (Nov/Dec 2017, Nov/Dec 2018, Apr/May 2018, Nov/Dec 2019 & Nov/Dec 2021) U

4. Explain in detail about the features of IR. (**Nov/Dec 2016 & Apr/May 2021) U**
5. Write short notes on (**Nov/Dec 2016)**
   i. Characterizing the web for search. **U**
   ii. Role of AI in IR. **AN**

6. Explain in detail about the Architecture of Information Retrieval. **(Apr/May 2022) U**
7. Explain in detail about the working of search engine. **(Apr/May 2021) U**
8. Analyze the challenges in IR system and give your suggestion to overcome that. **(Apr/May 2022) AN**
9. Brief about Open source search engine framework. **(Nov/Dec 2018 & Nov/Dec 2019) U**
10. Explain the impact of the web on Information retrieval systems. (**Nov/Dec 2018) AN**
11. How will you characterize the web? **U**
12. Write Short Notes on:
    i) The User Task **U**
    ii) IR Problem **AN**
    iii) Information and Data Retrieval. **U**
13. Explain in detail about Indexing, Retrieval and Ranking Process with a neat sketch. **U**
14. Explain the historical development of Information Systems. Discuss the sophistication in technology in detail. **U**

UNIT II MODELING AND RETRIEVAL EVALUATION

Basic IR Models – Boolean Model – TF-IDF (Term Frequency/Inverse Document Frequency) Weighting – Vector Model – Probabilistic Model – Latent Semantic Indexing Model – Neural Network Model – Retrieval Evaluation – Retrieval Metrics – Precision and Recall – Reference Collection – User-based Evaluation – Relevance Feedback and Query Expansion – Explicit Relevance Feedback.

Part A

1. **Define an Information Retrieval Model. (Apr/May 2022) R**

   Information Retrieval model specifies the representations that are used for documents and information needs, and how they are compared. It is a quadruple [D, Q, φ, R(qi, dj)]

   D -> representation for documents

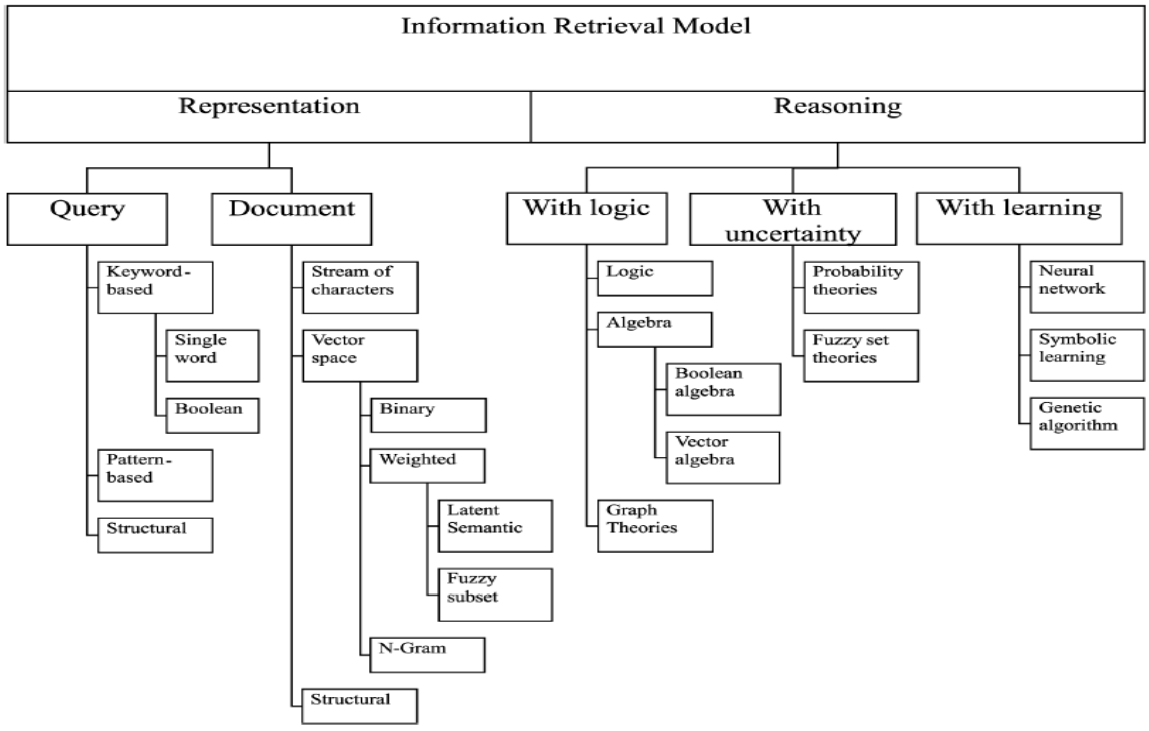
   Q -> representation for the user information (queries)

   Φ ->frame work for modeling document representations, queries and their relationships

   R(qi, dj) -> A ranking function which associates a real number with a query and document representation

2. **Give the taxonomy of Information Retrieval Model. (Apr/May 2022) R**

   The taxonomy consists of superimposing two views: vertical taxonomy, that classifies IR models with respect to a set of basic features, and horizontal taxonomy, which classifies IR systems and services with respect to the tasks they support.



3. **Can the tf-idf weight of a term in a document exceed 1? Why**? (**Apr/May 2018, Nov/Dec 2021) U**

YES, the tf-idf weight of a term in a document exceeds 1. TF-IDF is a family of measures for scoring a term with respect to a document (relevance). The simplest form of TF (word, document) is the number of times word appears in document. TFIDF can be 1 in the naive case, or to add the IDF effect, just do it log(number of documents/number of documents in which word is present).

4. **Consider the two texts," Tom and Jerry are friends" and "Jack and Tom are friends". What is the cosine similarity for these two texts? (Apr/May 2018, Nov/Dec 2021) U**

$$\cos\theta = \frac{a.b}{\sqrt{a^2}\sqrt{b^2}}$$

a : 1,1,1,1,1,0
b : 1,1,0,1,1,1
sim(a,b) = 1*1+1*1+1*0+1*1+1*1+0*1/Sqrt(5)*Sqrt(5)= 0.804

5. **What is Zone index**? **(Nov/Dec 2017) U**

A zone is a region of the document that can contain an arbitrary amount of text, e.g. Title, Abstract, References .We can build inverted indexes on zones as well to permit querying.

6. **State Bayes rule. (Nov/Dec 2017) U**

Bayes' theorem is stated mathematically as the following equation,

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

7. **What is politeness policies used in web crawling? (Nov/Dec 2017) U**

Web server has both implicit and explicit policies regulating the rate at which a crawler can visit them. Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site.

8. **What do you mean by relevance feedback? (Apr/May 2021 & Apr/May 2022) R**

Relevance feedback is a feature of some information retrieval systems. The idea behind relevance feedback is to take the results that are initially returned from a given query, to gather user feedback, and to use information about whether or not those results are relevant to perform a new query.

9. **List the retrieval models. (Nov/Dec 2016) U**
   a. **Boolean**
   b. **Vector space**

i. Basic vector space
ii. Extended Boolean model

c. **Probabilistic models**
i. Basic probabilistic model
ii. Bayesian inference networks
iii. Language models

d. **Citation analysis models**
i. Hubs & authorities (Kleinberg, IBM Clever)
ii. Page rank (Google)

10. What is cosine similarity? R

This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.

11. What are the disadvantages of Boolean model? U

The disadvantages of Boolean model are as follows:

- It is not simple to translate an information need into a Boolean expression
- Exact matching may lead to retrieval of too many documents.
- The retrieved documents are not ranked.
- The model does not use term weights.

12. Define term frequency. R

Term frequency: Frequency of occurrence of query keyword in document.

- More frequent terms in a document are more important, i.e. more indicative of the topic.
  f_{ij} = frequency of term i in document j
- May want to normalize *term frequency* (*tf*) by dividing by the frequency of the most common term in the document:
  $tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

13. What is inverse document frequency? R

Terms that appear in many *different* documents are *less* indicative of overall topic.

df_i = document frequency of term i
= number of documents containing term i
idf_i = inverse document frequency of term i,
= $\log_2 (N/ df_i)$
(N: total number of documents)

14. Describe the differences between vector space relevance feedback and probabilistic relevance feedback. (Nov/Dec 2020) AN

The tf-idf weighting is directly proportional to term frequency of the query term in the document whereas the probabilistic just takes into account the absence or presence of term in the document.

15. Define tf-idf weighting. (Apr/May 2019) U

The definitions of term frequency and inverse document frequency are combined to produce a composite weight for each term in each document. The *tf-idf* weighting scheme assigns to term *t* a weight in document *d* given by

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

In other words, $tf\text{-}idf_{t,d}$ assigns to term *t* a weight in document *d* that is

1. highest when *t* occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

16. Compare Term Frequency and Inverse Document Frequency. (Apr/May 2021) AN

Term frequency refers to the number of times that a term t occurs in document d. The inverse document frequency is a measure of whether a term is common or rare in a given document corpus. It is obtained by dividing the total number of documents by the number of documents containing the term in the corpus.

17. List the two key statistics that are used to assess the effectiveness of an IR system. U

- Precision
- Recall

18. Define the term Stemming. (Nov/Dec 2018) R

Conflation algorithms are used in information retrieval systems for matching the morphological variants of terms for efficient indexing and faster retrieval operations. The Conflation process can be done either manually or automatically. The automatic conflation operation is also called stemming.

19. What is Recall? R

Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents retrieved.

20. What is precision? R

Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

21. Define Latent semantic Indexing. U

Latent Semantic Indexing is a technique that projects queries and documents into a space with "latent" Semantic dimensions. It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden.

22. Is the vector space model always superior to the Boolean model? AN

No. Information retrieval using the Boolean model is usually faster than using the vector space model. I believe that Boolean retrieval is a special case of the vector space model, so if you look at ranking accuracy only, the vector space gives better or equivalent results.

23. What are the advantages and limitations of the vector space model? U

Advantages:

1.It is a simple model based on linear algebra

2.There weights are not binary

3. Allows the computing for a continuous degree of similarities between queries and documents.

Disadvantages:

1. Suffers from synonym and polysemy

2. It theoretically assumes that terms are statistically independent.

24. What are the advantages and disadvantages of TF-IDF? U

Advantages:

- Easy to compute
- Basic metric to extract the most descriptive terms in a document
- Easily compute the similarity between 2 documents using it

Disadvantages:

- TF-IDF is based on the bag-of-words (BoW) model, therefore it does not capture position in text, semantics, co-occurrences in different documents, etc.
- Cannot capture semantics (e.g. as compared to topic models, word embedding)

25. What's the difference between TF-IDF and DF-ICF? U

Tf-idf is the weight given to a particular term within a document and it is proportional to the importance of the term.

- Tf-idf = Term Frequency - inverse document frequency;
- df-icf = Document Frequency - inverse corpus frequency.

The first is weighted importance of the term (or word) in the document; the second is the weighted importance of the document in the corpus. Treat each tweet as a document, and the words within it as terms.

26. Is information retrieval different from information extraction? If yes, how? AN

Information Extraction involves functional and structural execution of input requirement. On the other hand, Information Retrieval is getting relatively larger structures of data relevant to specific 'query' (mostly string queries , with relevance determined by proportion of fulfillment of queries made on huge collections of 'documents' of data.

27. Define Language model. U

A function that puts a probability measure over strings which drawn from some vocabulary**.**

$$\sum_{s \in \Sigma^*} P(s) = 1$$

28. List the applications of language models. U

- Speech Recognition
- Spelling Correction
- Machine Translation

29. Is it necessary to do query expansion always? Why? (Apr/May 2019) AN

As formulating well-designed queries is difficult for most users, it is necessary to use query expansion to retrieve relevant information. Query expansion techniques are widely applied for improving the efficiency of the textual information retrieval systems.

30. Define Query Expansion. U

Query expansion techniques are usually based on an analysis of word or term co-occurrence, in either document collection, a large collection of queried or the top ranked documents in a result list.

Part B & C

1. Write short notes on the following: **(Nov/Dec 2017) U**
   a. Probabilistic relevance feedback
   b. Pseudo relevance feedback
   c. Indirect relevance feedback
2. Explain in detail about binary independence model for Probability Ranking Principle (PRP). (**Nov/Dec 2017**, **Nov/Dec 2018) U**
3. Write short notes on Latent semantic Indexing (LSI). **(Nov/Dec 2018, Apr/May 2018, Nov/Dec 2019, Apr/May 2019 & Nov/Dec 2020) U**
4. Discuss the query likelihood model in detail and describe the approach for information retrieval using this model. **(Apr/May 2018) U**
5. How we do process a query using an inverted index and the basic Boolean Retrieval model? **Nov/Dec 2018, Nov/Dec 2019 & Apr/May 2019) U**
6. Describe how the query generation probability for query likelihood model. **(Nov/Dec 2017) U**
7. Briefly explain weighting and Cosine similarity. **(Nov/Dec 2016 , Nov/Dec 2021) U**
8. Write about relevance feedback and query expansion. **(Nov/Dec 2016 , Nov/Dec 2021) U**
9. Compare and Contrast Boolean, Vector Space and Probabilistic model in detail. **(Apr/May 2022) AN**

10. Explain in detail about vector space model for documents. **(Nov/Dec 2018 & Apr/May 2022)** **U**
11. Explain TF-IDF weighting. **(Apr/May 2022)** **U**
12. Describe processing with sparse vectors. **U**
13. Explain probabilistic IR. **U**
14. Describe the language model based IR. **(Nov/Dec 2020)** **U**
15. When does relevance feedback work? **(Nov/Dec 2018, Nov/Dec 2020)** **AN**

UNIT III TEXT CLASSIFICATION AND CLUSTERING

A Characterization of Text Classification – Unsupervised Algorithms: Clustering – Naïve Text Classification – Supervised Algorithms – Decision Tree – k-NN Classifier – SVM Classifier – Feature Selection or Dimensionality Reduction – Evaluation metrics – Accuracy and Error – Organizing the classes – Indexing and Searching – Inverted Indexes – Sequential Searching – Multi-dimensional Indexing.

Part A

1. List the basic types of machine leaning algorithm, depending on the learning mechanisms used. R

Depending on the learning mechanisms used, the machine learning algorithm can be basically of 3 types

- Supervised leaning
- Unsupervised learning
- Semi-supervised learning

Other types of learning algorithm includes

- Reinforcement algorithm and
- Transduction

2. Differentiate between relevance feedback and pseudo relevance feedback. (Nov/Dec 2018) U

The idea of relevance feedback ( ) is to involve the user in the retrieval process so as to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results.

Pseudo relevance feedback, also known as blind relevance feedback, provides a method for automatic local analysis. It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction.

3. Discuss about Curse of Dimensionality. (Apr/May 2022) U

The curse of dimensionality basically means that the error increases with the increase in the number of features. It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions. It's easy to catch a caterpillar moving in a tube (1 dimension). It's harder to catch a dog if it were running around on the plane (two dimensions). It's much harder to hunt birds, which now have an extra dimension they can move in.

4. Define Text Categorization. (Apr/May 2022) U

Text categorization (text classification) is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world.

For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively.

5. How do spammers use cloaking to server spam to the web users? (Apr/May 2018) U

Cloaking is a spamming technique in which the content presented to the search engine spider is different from the content presented to regular users. This is done by delivering content based on the user-agent HTTP header of the user requesting the page, the IP address of a user or the referring page: A web page can be cloaked based on the IP address, the user-agent, referring web page or any combination of these three factors.

6. What is hierarchical agglomerative clustering? (Apr/May 2021) U

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents.

7. Can a digest of the characters in a web page be used detect near duplicate web pages? Why? (Apr/May 2018) U

For every single web page, calculating a fingerprint that is a succinct (say 64-bit) digest of the characters on that webpage is the modest method for detecting duplicates. When the fingerprints of two webpage documents are the same, at that point we have to examine if the pages are the same and if so, and then state that one of those to be a duplicate copy of the other.

8. What is inversion in indexing process? (Nov/Dec 2017, Apr/May 2021) R

The core of indexing process converts document –term information to term document for indexing.

9. Define Text clustering. U

Given a collection D of documents, a text clustering method automatically separates these documents into K clusters according to some predefined criteria.

10. Define Document Preprocessing. (Nov/Dec 2016) R

It is a collection of activities in which Text Documents are pre-processed. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help Text mining such as prepositions, articles, and pro-nouns can be eliminated. The preprocessing includes,

- Lexical Analysis
- Elimination of Stop words
- Stemming

- Index Terms Selection
- Thesaurus

11. What are the requirements of XML information retrieval systems?
(Nov/Dec 2016 & Nov/Dec 2021) U

The requirements are,

- markup of logical structure of documents
- separation of logical structure and layout
- support interoperability

12. What are the parameters to identify spam? (Apr/May 2019) AN

- It has a broadcasted, rather than targeted, message
- It suits the purposes of the sender rather than the receiver
- Most important, the message is distributed without the explicit permission of the recipients

Many messages that constitute spam in the minds of list users might not be intended as spam by the sender, but spam is in the mind's eye of the receiver.

13. What is SVM? U

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.

SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

14. Describe benefit of SEO. U

- Increase your search engine visibility
- Generate more traffic from the major search engines
- Make sure your website and business get NOTICED and VISITED
- Grow your client base and increase business revenue

15. Explain the difference between SEO and Pay-per-click. AN

SEO	Pay-Per-click
SEO results take 2 weeks to 4 months	It results in 1-2 days
It is very difficult to control flow of traffic	It has ability to turn on and at any moment
Requires ongoing learning and experience to reap results	Easier for a novice
It is more difficult to target local markets	Ability to target "local" markets
Better for long-term and lower margin campaigns	Better for short-term and high-margin campaigns.
Generally more cost-effective , does not penalize for more traffic	Generally more costly per visitor and per conversion

16. What is the Near-duplicate detection? R

Near-duplicate is the task of identifying documents with almost identical content. Near- duplicate web documents are abundant. Two such documents differ from each other in a very small portion that displays advertisements, for example. Such differences are irrelevant and for web search.

17. List the advantages of block addressing. R

Block addressing allows the pointer to be smaller, because there are fewer blocks than positions. Also, all the occurrences of a word inside a single block are collapsed to one reference. Indexes of only 5% overhead over the text size are obtained with this technique.

18. Differentiate text centric Vs. data centric XML retrieval. AN

XML structure serves as a framework within which we match the text of the query with the text of the XML documents. This exemplifies a system that is optimized for text-centric XML. While both text and structure are important, we give higher priority to text

The XML document retrieval is characterized by (i) long text fields (e.g., sections of a document), (ii) inexact matching, and (iii) relevance-ranked results.

In contrast, data-centric XML mainly encodes numerical and non-text attribute-value data. When querying data-centric XML, we want to impose exact match conditions in most cases. This puts the emphasis on the structural aspects of XML documents and queries. An example is: Find employees whose salary is the same this month as it was 12 months ago.

19. Define XML retrieval. R

XML retrieval, or XML Information Retrieval, is the content-based retrieval of documents structured with XML (eXtensible Markup Language). As such it is used for computing relevance of XML documents.

20. Define spam. (Nov/Dec 2019) R

Spam is usually considered to be electronic junk mail or junk newsgroup postings. Some people define spam even more generally as any unsolicited email.

Part B & C

1. Explain in detail about naïve bayes algorithm and its application in text classification.

(Nov/Dec 2019, Nov/Dec 2020, Apr/May 2021 & Apr/May 2022) U

2. Discuss in detail about SVM classifier and their use in text classification. **(Apr/May 2022) U**

3. Explain K-Means clustering in grouping different documents.

(Nov/Dec 2019 & Nov/Dec 2020) U

4. Elaborate K-Nearest Neighbor algorithm with an illustration. **(Apr/May 2021) U**

5. You are asked to design a text classification engine to process all queries raised by the employees

of your organization. Elaborate in detail about the steps you will take and various factors to be considered while designing. **(Apr/May 2021) C**

6. Describe the distributing indexes. **U**
7. Explain the paid placement. **U**
8. Explain web indexes. **U**
9. Explain the basic XML concepts. **U**
10. Illustrate the various challenges in XML retrieval with appropriate examples. **AN**
11. Explain the Rocchio algorithm for relevance feedback. **(Nov/Dec 2020 & Nov/Dec 2021) U**
12. Explain in detail about vector space model for XML retrieval.

(Nov/Dec 2019, Apr/May 2019, Nov/Dec 2020 & Nov/Dec 2021) U

UNIT IV WEB RETRIEVAL AND WEB CRAWLING

The Web – Search Engine Architectures – Cluster based Architecture – Distributed Architectures – Search Engine Ranking – Link based Ranking – Simple Ranking Functions – Learning to Rank – Evaluations - Search Engine Ranking – Search Engine User Interaction – Browsing – Applications of a Web Crawler – Taxonomy – Architecture and Implementation – Scheduling Algorithms – Evaluation.

PART A

1. What is peer-to peer search? (Nov/Dec 2017, Nov/Dec 2018) R

A distributed search engine is a search engine where there is no central server. Unlike traditional centralized search engines, work such as crawling, data mining, indexing, and query processing is distributed among several peers in a decentralized manner where there is no single point of control.

2. What are the performance measures of search engine? (Nov/Dec 2017, Nov/Dec 2018) U

- Speed of response /size of index are factors
- Need a way of quantifying user happy.
- Precision, Recall
- Technical precision
- Pay per method

3. What is snippet generation? (Nov/Dec 2017 & Nov/Dec 2021) R

A document retrieval system generates snippets of documents for display as part of a user interface screen with search results. The snippet may be generated based on the type of query or the location of the query terms in the document. It's a short summary of the document which allows the user to decide its relevance.

4. What are the applications of web crawlers? (Nov/Dec 2018) U

- Web Search engine
- Web Archiving
- Web Data mining
- Web Monitoring
- Web scraping
- Web Mirroring

5. Define Meta Crawler. (Nov/Dec 2020) R

The Meta Crawler is a free search service used for locating information available on the World Wide Web. The Meta Crawler has an interface similar to WebCrawler and Open Text in that it

allows users to enter a search string, or query, and returns a page with click-able references or hits to pages available on the Web.

6. What is the purpose of web crawler? (Nov/Dec 2016, Nov/Dec 2021 & Apr/May 2021) U

A Web crawler, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of indexing. The purpose of web crawler includes,

- Creating a Localized Search Engine.
- Load Testing from multiple server locations & different countries.
- Detecting SEO Optimizations on Various Pages, like missing Meta tags.
- Generating Customized Reports, which log file analysis tools, might not create.
- Spell-Checking Pages when working on large sites

7. Define Search Engine Optimization. (Nov/Dec 2018) R

Search Engine Optimization is the act of modifying a website to increase its ranking in organic, crawler-based listing of search engines. There are several ways to increase the visibility of your website through the major search engines on the internet today.

8. List the characteristics of Map Reduce Strategy. (Nov/Dec 2017 & Nov/Dec 2021) R

- Very large scale data: peta, Exabyte
- Write once and read many data
- Map and reduce operation are typically performed by same physical processor.
- Number of map tasks and reduce tasks are configurable.
- Operations are provisioned near the data.
- Commodity hardware and storage.

9. Define focused crawler. R

A focused crawler or topical crawler is a web crawler that attempts to download only pages that are relevant to a pre-defined topic or set of topic.

10. What is hard and soft focused crawling? R

In **hard focused crawling** the classifier is invoked on a newly crawled document in a standard manner. When it returns the best matching category path, the out-neighbors of the page are checked into the database if and only if some node on the best matching category path is marked as good.
In **soft focused crawling** all out-neighbors of a visited page are checked into DB2, but their crawl priority is based on the relevance of the current page.

11. Mention of the features of a crawler. U

1. Robustness
2. Politeness
3. Distributed
4. Scalable
5. Performance and Efficiency

6. Quality
7. Freshness
8. Extensible

12. Define Authorities. (Nov/Dec 2016) R

Authorities are pages that are recognized as providing significant, trustworthy and useful information on a topic. In-degree is one simple measure of authority. However in-degree treats all links as equal.

13. Define hubs. R

Hubs are index pages that provide lots of useful links to relevant content pages. Hub pages for IR are included in the home page.

14. What is Hadoop? R

The Map-Reduce operation run on a special file system called Google File System that is highly optimized for this purpose. GFS is not open source. Doug Cutting and Yahoo! reverse engineered the GFS and called it Hadoop Distributed File System. The software framework that supports HDFS, Map Reduce and other related entities is called the project Hadoop or simply Hadoop.

15. What is the Hadoop Distributed File System? R

The Hadoop Distributed File System is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user application. HDFS stores file system metadata and application data separately. The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes, which record attributes like permissions, modification and access times, namespace and disk space quatas.

16. Define Map-Reduce. R

Map-Reduce is a programming model and software framework first developed by Google.
Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

17. How Cross-Lingual Retrieval is typically implemented?

(Apr/May 2018 & Nov/Dec 2020) R

Cross – Lingual Retrieval refers to the retrieval of documents that are in a language different from the one in which the query is expressed. This allows users to search document, collections in multiple language and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages.

18. What is collaborative filtering? (Apr/May 2019) R

Collaborative filtering is a method of making automatic predictions about the interests of a single user by collecting preferences or taste information from many users. It uses given rating data by many users for many items as the basic for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user.

19. Define user based collaborative filtering. (Nov/Dec 2016) R

User-based CF algorithm produces recommendation list for object user according to the view of other users. The assumptions are if the ratings of some items rated by some users are similar, the rating of other items rated by these users will also be similar.

20. What do you mean by item-based collaborative filtering? (Nov/Dec 2018) R

Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.

21. What is a page rank? R

PageRank (PR) is a quality metric invented by Google's owners Larry Page and Sergey Brin. The values 0 to 10 determine a page's importance, reliability and authority on the web according to Google.

22. How do you compute PageRank values? (Nov/Dec 2020) AP

PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. What that means to us is that we can just go ahead and calculate a page's PR without knowing the final value of the PR of the other pages. That seems strange but, basically, each time we run the calculation we're getting a closer estimate of the final value. So all we need to do is remember the each value we calculate and repeat the calculations lots of times until the numbers stop changing much.

23. Define hub score and authority score. R

For any query, we compute two ranked lists of results rather than one. The ranking of one list is induced by the hub scores and that of the other by the authority scores.

24. How to handle invisible web? (Nov/Dec 2018) AN

The internet is an iceberg. And, as you might guess, most of us only reckon with the tip. While the pages and media found via simple searches may seem unendingly huge at times, what is submerged and largely unseen – often referred to as the invisible web or deep web – is in fact far, far bigger. A number of subject-specific databases, engine and tools with an established filter, making their searches much narrower. Open Access Journal Databases can be used to handle invisible web.

25. What is good clustering? (Nov/Dec 2018) R

A good clustering method will produce high quality clusters in which the quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

26. Compute the jaccard's similarity for the two list of words (time, flies, like, an, arrow) and (how, time, flies). (Apr/May 2018) U

The *Jaccard similarity* is defined

$$\mathrm{JS}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A:time,flies,like,an,arrow

B:how,time,flies

A	B
1	1
1	1
1	0
1	0
1	0
0	1

JS(A,B)=2/6=0.804

27. What is the difference between Clustering and Collaborative Filtering? U

Clustering is structure-finding, typically among dense data of low or moderate dimension in a continuous space. It's really defined by a distance function among data points. It typically employs some form of expectation-maximization-style algorithm.

Collaborative filtering is in general a ranking problem. Depending on how you look at it, the data are sparse, high-dimensional and in a continuous space. It amounts to inferring which missing dimension has the highest value. It typically proceeds via a matrix completion algorithm like low-rank factorization.

28. What is HITS algorithm? U

Hypertext Induced Topic selection (HITS) (**HITS**; also known as **hubs and authorities**) is a link analysis algorithm that rates Web pages. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages.

29. Define politeness with respect to a crawler. U

Crawlers should also fulfill politeness

- A crawler cannot overload a Web site with HTTP requests
- A crawler should wait a small delay between two requests to the same Web site

30. Outline the differences between web search and information retrieval. (Apr/May 2022) AN

Information retrieval systems (IRS) are field concerned with retrieval of information. A search engine is the application of IR techniques. A web search engine is a tool to find information on the www. Search engines are updating their index to the World Wide Web.

31. Define recall and precision in the context of a search engine. (Apr/May 2022) U

Recall is the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

32. How does relevance scoring works in web search? (Apr/May 2019) U

Search relevance is the measure of accuracy of the relationship between the search query and the search results. Online users have high expectations. Thanks to the high bar set by sites like Google, Amazon, and Netflix, they expect accurate, relevant, and rapid results.

33. What is cross lingual retrieval? (Nov/Dec 2019) U

Cross-language information retrieval (**CLIR**) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query.[1] The term "cross-language information retrieval" has many synonyms, of which the following are perhaps the most frequent: cross-lingual information retrieval, trans lingual information retrieval, multilingual information retrieval.

Part B & C

1. Consider a web graph with three nodes 1, 2 and 3.The links are as follows: 1→ 2, 3 → 2, 2→1, 2→3.Write down the transition probability matrices for the surfers walk with teleporting, for the teleport probability: a=0.5 and compute the page rank. **(Nov/Dec 2018 & Nov/Dec 2021) C**
2. Explain in detail about Community-based Question Answering system.

   (**Nov/Dec 2017, Nov/Dec 2018, Nov/Dec 2021) U**
3. How do the various nodes of a distributed crawler communicate and share URLs? (**Nov/Dec 2018 & Nov/Dec 2021) U**
4. Explain in detail about finger print algorithm for near-duplicate detection.

   (Nov/Dec 2017, Nov/Dec 2018, Nov/Dec 2021 & Apr/May 2021) U
5. Brief about search engine optimization.

   (Nov/Dec 2017, Nov/Dec 2020, Nov/Dec 2021 & Apr/May 2021) U
6. Elaborate on the search engine architectures.

   (Nov/Dec 2016, Nov/Dec 2021, Apr/May 2021 & Apr/May 2022) U
7. Describe meta and focused crawling. **(Nov/Dec 2016) U**
8. Explain the features and architecture of web crawler.

   (Nov/Dec 2018, Apr/May 2018, Apr/May 2019 & Nov/Dec 2021) U
9. Explain about online selection in web crawling. **(Nov/Dec 2018 & Nov/Dec 2021) U**

10. Discuss the design of a Question–Answer engine with the various phases involved. How can the performance of such an engine be measured? **(Apr/May 2018 & Apr/May 2019) R**
11. Brief on Personalized search. (**Nov/Dec 2017, Nov/Dec 2018 & Nov/Dec 2021) R**
12. Explain in detail, the Collaborative Filtering using clustering technique.
    (**Nov/Dec 2017 & Nov/Dec 2021) R**
13. Brief about HITS link analysis algorithm.
    (Nov/Dec 2016, Nov/Dec 2017, Nov/Dec 2018, Nov/Dec 2019, Apr/May 2019, Nov/Dec 2020, Nov/Dec 2021 & Apr/May 2022) R
14. Explain in detail cross lingual information retrieval and its limitations in web search.
    (Nov/Dec 2016) U
15. How does Map reduce work? Illustrate the usage of map reduce programming model in Hadoop. **(Apr/May 2018, Nov/Dec 2019, Apr/May 2019 & Nov/Dec 2020) U**
16. Explain the page rank computation. **AP**
17. Describe Markov chain process. **U**
18. Explain web as a graph. **U**
19. Explain relevance scoring. **U**
20. How to handle invisible web? **AN**
21. Discuss about the snippet generation. **(Apr/May 2019) U**
22. Explain summarization. **(Apr/May 2019) U**
23. Discuss in detail about the typical user interaction models for the most popular Search Engines of today. **U**
24. How is Crawling evaluation done? **AN**

UNIT V RECOMMENDER SYSTEM

Recommender Systems Functions – Data and Knowledge Sources – Recommendation Techniques – Basics of Content-based Recommender Systems – High Level Architecture – Advantages and Drawbacks of Content-based Filtering – Collaborative Filtering – Matrix factorization models – Neighborhood models.

Part A

1. **Differentiate supervised and unsupervised learning. (Nov/Dec 2017) AN**
   In supervised learning, the output datasets are provided which are used to train the machine and get the desired outputs whereas in unsupervised learning no datasets are provided, instead the data is clustered into different classes.

2. **What is Dendrogram? (Nov/Dec 2017) R**
   Decompose data objects into a several levels of nested partitioning called a dendrogram. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

3. **Give an example for a Dendrogram**. (**Apr/May 2018) U**
   - Tiered Diagram used to display the playoff games and progress of some sporting event like hockey, basketball or baseball.
   - Agglomerative hierarchical clustering
   - Divisive Hierarchical clustering

4. **What is the use of the Expectation-Maximization algorithm? (Apr/May 2018) U**
   The **Expectation Maximization (EM)** algorithm can be used to generate the best hypothesis for the distributional parameters of some multi-modal data. The EM algorithm can be used to estimate latent variables, like ones that come from mixture distributions (you know they came from a mixture, but not which specific distribution).

5. **What are the characteristics of information filtering? (Nov/Dec 2016) U**
   - Document set : Dynamic
   - Information need : Stable, long term, specified in a user profile
   - Profile :Highly personalized
   - Selection process : Delegated

6. What are the desirable properties of a clustering algorithm?

(Nov/Dec 2016 & Nov/Dec 2021) U

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incorporation of user-specified constraints
- Interpretability and usability

7. What do you mean by information filtering? (Nov/Dec 2021) R

An information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or computerized methods prior to presentation overload and increment of the semantic signal-to-noise ratio.

8. Explain difference between information filtering and information Retrieval.

(Nov/Dec 2018 & Nov/Dec 2020) AN

Information Filter	Information Retrieval
IF is concerned with the removal of textual information from an incoming stream and its dissemination to groups or individuals.	IR systems are concerned with the collection and organization of texts so that users can then easily find a text in the collection.
Information filtering is concerned with repeated uses of the system by users with long-term, but changing interests and needs.	A query represents a one-time information need.
Filtering is based on descriptions of individual or group interests or needs that are usually called profiles.	Retrieval of information is instead based on user specified information needs in the form of a query.
IF systems deal with dynamic data.	IR systems deal with static databases.

9. What is text mining? (Apr/May 2019) R

Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories.

Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

10. What is classification? R

Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”.

11. What is decision tree? R

- A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. A decision tree or a classification tree is a tree in which each internal node is labeled with an input features.
- The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

12. List the advantages of decision tree. U

- Decision tree can handle both nominal and numeric input attributes.
- Decision tree representation is rich enough to represent any discrete value classifier.
- Decision trees are3 capable of handling database that may have errors.
- Decision trees are capable of handling datasets that may have missing values.
- It is self-explanatory and when compacted they are also easy to follow.

13. List the disadvantages of decision tree. U

- Most of the algorithms require that the target attribute will have only discrete values.
- Most decision-tree algorithms only examine a single field at a time.
- Decision trees are prone to errors in classification problems with much class.
- As decision tree use the "divide and conquer" method, they tend to perform well if a few highly relevant attribute exists, but less so if many complex interactions are present.

14. State the applications of clustering in information retrieval. AN

Application	What is clustered?	Benefit
Search result clustering	search results	more effective information presentation to user
Scatter-Gather	(subsets of) collection	alternative user interface: "search without typing"
Collection clustering	collection	effective information presentation for exploratory browsing
Language modeling	collection	increased precision and/or recall
Cluster-based retrieval	collection	higher efficiency: faster search

15. What is a recommender system? Explain. (Apr/May 2022) U

A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that provide suggestions for items that are most pertinent to a particular user. Loosely defined, a recommender system is a system which predicts ratings a user might give to a specific item. These predictions will then be ranked and returned back to the user. They're used by various large name companies like Google, Instagram, Spotify, Amazon, Reddit, Netflix etc.

16. What is meant by text preprocessing? (Nov/Dec 2020) U

Text preprocessing involves transforming text into a clean and consistent format that can then be fed into a model for further analysis and learning. Text preprocessing techniques may be general so that they are applicable to many types of applications, or they can be specialized for a specific task.

17. Write the pros and cons of using classification in text mining over clustering algorithms. (Apr/May 2019) AN

There is a fundamental difference between classification, which is abstract and definitive, and grouping, which depends on the characteristics of a particular sample or set of observations that is concrete and is thus descriptive.

Advantages of Classification:

It facilitates the identification of organisms. It explains how different creatures interact with one another. It aids in the comprehension of organism evolution. It helps to understand how animals, plants, and other living creatures are related and how they can benefit humans.

Disadvantages of Classification:

A disadvantage to classification is that many of the classifications themselves are based on subjective judgments, which may or may not be shared by everyone participating.

18. Outline the difference between classification and clustering. (Nov/Dec 2019) AN

Classification

- It is used with supervised learning.
- It is a process where the input instances are classified based on their respective class labels.
- It has labels hence there is a need to train and test the dataset to verify the model.
- It is more complex in comparison to clustering.
- Examples: Logistic regression, Naive Bayes classifier, Support vector machines.

Clustering

- It is used with unsupervised learning.
- It groups the instances based on how similar they are, without using class labels.
- It is not needed to train and test the dataset.
- It is less complex in comparison to classification.
- Examples: k-means clustering algorithm, Gaussian (EM) clustering algorithm.

19. State Bayes theorem. (Nov/Dec 2019) U

Bayes theorem gives the probability of an "event" with the given information on "tests". There is a difference between "events" and "tests". For example there is a test for liver disease, which is different from actually having the liver disease, i.e. an event. Rare events might be having a higher false positive rate.

Part B & C

1. Explain in detail about the different classes of recommendation approaches along with its advantages and disadvantages. **(Apr/May 2021) U**
2. With respect Recommendation system explain in detail about the data and knowledge sources. **U**
3. Explain in detail about the high-level architecture of content-based system. **U**
4. Enlist the eleven popular tasks proposed by Herlocker that a RS can assist in implementation. **U**
5. Why Product Recommendation Engines Are Not Good Product Search Engines? Explain **U**
6. Explain the role of Matrix factorization techniques in Collaborative filtering. **(Apr/May 2021) U**
7. Explain in detail about Neighborhood models of Collaborative filtering. **(Apr/May 2022) U**
8. Compare the accuracy of collaborative filtering algorithms using Netflix data. **AN**
9. Write short note on text mining. **(Nov/Dec 2021) R**
10. Explain in detail about agglomerative clustering. Compare it with other clustering algorithms. **(Nov/Dec 2021) U**
11. Explain decision tree algorithm with example. **(Nov/Dec 2019, Apr/May 2019 & Nov/Dec 2020) R**
12. Explain K-means algorithm of clustering with example. **(Nov/Dec 2019 & Nov/Dec 2020) R**
13. Explain the application of Expectation-Maximization (EM) algorithm in text mining. **(Apr/May 2021) AN**