UNIT I – INTRODUCTION

Part A – Question Bank

1. **Define information retrieval.**
   Information Retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

2. **Explain difference between data retrieval and information retrieval.**

Parameters	Data Retrieval	Information retrieval
Example	Data Base Query	WWW Search
Matching	Exact	Partial Match, Best Match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic

3. **List and explain components of IR block diagram.**
   - **Input** – Store Only a representation of the document
   - **A document representative** – Could be list of extracted words considered to be significant.
   - **Processor** – Involve in performance of actual retrieval function
   - **Feedback** – Improve
   - **Output** – A set document numbers.

4. **What is objective term and nonobjective term?**
   Objective Terms – Are extrinsic to semantic content, and there is generally no disagreement about how to assign them.
   Nonobjective Terms – Are intended to reflect the information manifested in the document, and there is no agreement about the choice or degree of applicability of these terms.

5. Explain the type of natural language technology used in information retrieval.

Two types

I. Natural language interface make the task of communicating with the information source easier, allowing a system to respond to a range of inputs.

II. Natural Language text processing allows a system to scan the source texts, either to retrieve particular information or to derive knowledge structures that may be used in accessing information from the texts.

6. What is search engine?

A search engine is a document retrieval system design to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.

7. What is conflation?

Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. The process of stemming if often called conflation.

8. What is an invisible web?

Many dynamically generated sites are not index able by search engines; This phenomenon is known as the invisible web.

9. Define Zipf's law.

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

10. What is open source software?

Open source software is software whose source code is available for modification or enhancement by anyone.

"Source code" is the part of software that most computer users don't ever see; it's the code computer programmers can manipulate to change how a piece of software—a "program" or "application"—works. Programmers who have access to a computer program's source code can improve that program by adding features to it or fixing parts that don't always work correctly.

11. What is proprietary software?

Proprietary software is computer software which is the legal property of one party. The term of use for other parties is defined by contracts or licensing agreements.

These terms may include various privileges to share, alter , dissemble, and use the software and its code.

12. What is closed software?

Closed software is a term for software whose license does not allow for the release or distribution of the software's source code. Generally it means only the binaries of a computer program are distributed and the license provides no access to the programs source code. The source code of such programs is usually regarded as a trade secret of the company. Access to source code by third parties commonly requires the party to sign a non-disclosure agreement.

13. List the advantage of open source.

- 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.

14. List the disadvantage of open source.

- 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.

15. What are the reasons for selecting open software?

- Development and maintenance of open source software is a community based activity.
- Open source software licenses are copyright protected they strictly ensure the user freedom to use, modify and distribute the programs.
- Is interoperable customizable according to the needs and fulfills the software industry standards.
- Open source software allows everyone to use, study, modify and distribute the software.
- Allows a broader perspective when comes to its support.

16. What do you mean by Apache License?

- The Apache License is a free software license written by the Apache Software Foundation (ASF). The name Apache is a registered trademark and may only be used with the trademark holders express permission.

- Apache license is a high performance, Full-featured text search engine library written entirely in Java.

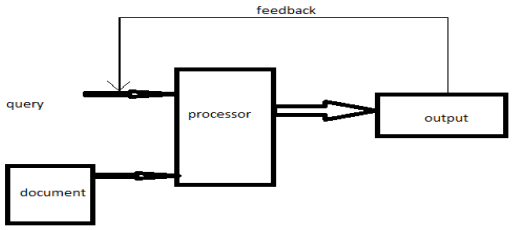
17. Explain features of GPL version2.

- It gives permission to copy and distribute the programs unmodified source code.
- It allows modifying the programs source code and distributing the modified source code.
- User distributes compiled versions of the program, both modified and unmodified.
- All modified copies are distributed under the GPL v2.
- All compiled versions of the program are accompanied by the relevant source code.

Part –B

1. Explain about IR Architecture.
2. **Components of information retrieval**:

Information retrieval system is an information system, which is used to store items of information that need to be processed, searched, disseminated and retrieved to various user populations. It consists of three components: input, processor, and output.



a) **Input**: store only representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.

b) A **document representative** could be a list of extracted words considered to be significant.

c) **Processor**: involve in performing actual retrieval function, executing the search strategy in response to a query.

d) **Feedback**: Improving the subsequent run after sample retrieval.

e) **Output:**A set of document numbers.

3. Explain about Components of search engine..

Search enginesare becoming the primary entry point for discovering the web pages. Ranking of web pages will influences which pages user will view. Exclusion of a site from search engines will cut off the site from its intended audience. The privacy policy of search engine is important.
Most successful search engines use centralized architecture and global ranking algorithms to generate the ranking of documents crawled in the databases, for example, Google's PageRank.

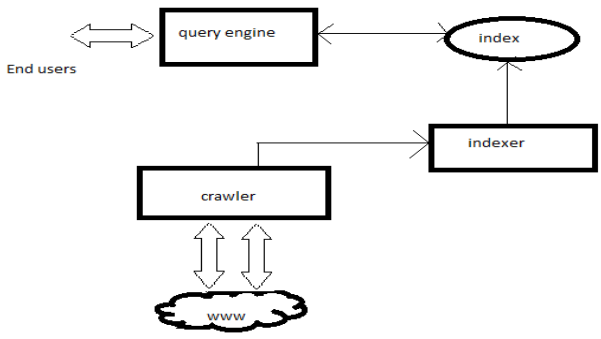
- Centralized Architecture
- Distributed Architecture

Centralized Architecture:

Using crawlers, information is gathered into a single site, where it is indexed; the site processes all the queries.

Centralized architecture consists of following components:

1. Crawlers
2. Index
3. Query engine
4. User interface.



Problems using this architecture:

- Difficult o gathering the data because of dynamic nature of the web.
- The Communication problem.

Distributed architecture:

It doesn't have its own actual record database. It just indexes the interfaces of sub database system. Example, Harvest.

Harvest: it is used to save bandwidth by deploying gatherers near the data source and exchanging the summarized data which usually is much smaller than the original data.
Architecture:

4. Discuss about open source.

open software:

Open source software is like any other software. This software is differentiating by its use and licenses. Open source software guaranteesthe right to access and modify the source code and to use reuses and redistribute the software, all with no royalty or other costs.

Need:

It is fact that single solution provider can not produce all the needed solutions. Open software are now available for anyone and any use.

Free software foundation (FSF)
Open source
Proprietary software
Closed source

Success of open source:
Operating systems
Servers
Programming languages
Client software
Digital content
Open source software
Open standards

Advantages:

1. The right to use the software in any way.
2. There is usually no license cost and free of cost.
3. The source code is open and can be modified freely
4. Possible to reuse the source code.
5. Open standards.
6. It provides higher flexibility.

Disadvantages:

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