General Questions

1 – What is the definition of Information retrieval?

Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

2 – What is the definition of Data Extraction?

Is a process that involves retrieval of data from various sources.

3 – Why companies extract data?

Companies extract data in order to:

- **Process it.**
- **Analyze it.**
- **Migrate the data to a data repository.**

4 – What is the definition of Information Extraction?

Is the automated retrieval of specific information related to a selected topic from a body or bodies of text.

5 – Mention one usage for Information Extraction Tools?

Information extraction tools make it possible to pull information from text documents, databases, websites or multiple sources.

6 – What is the definition of Data Mining?

Is the method of analyzing expansive sums of data in an exertion to discover relationships, designs, and insights.

7 – The Designs of data in Data Mining should be.....

meaningful in that they lead to a few advantages "financial advantage".

8 – What is the definition of Web Mining?

- Is the method of utilizing data mining strategies and calculations to extract information specifically from the net by extricating it from web documents and services, substance, hyperlinks and server logs.

- It also includes the method of finding valuable and obscure data from web information.

9– What is the objective of Web Mining?

search for the designs in web information by collecting and analyzing data in order to urge insights.

10 – What is the definition of Web Crawler or Spider?

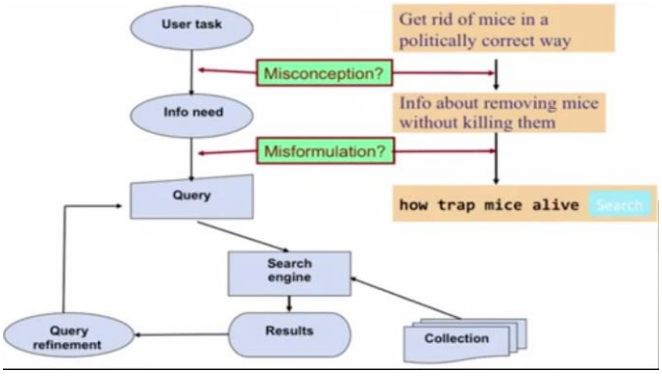
Is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.

11 – What is the definition of Web Scraper?

Is a process of extracting specific data. Unlike web crawling, a web scraper searches for specific information on specific websites or pages.

12 – What is the Basic Assumptions of Information Retrieval (With Diagram)?

- **Collection:** A set of documents.
- **Goal:** retrieve documents with Information that is relevant to the user's information need and help the user to complete a task.



13 – Define the term "Information Need"?

An information need is the topic about which the user desires to know more.

14 –What's the different between Information need and query?

It is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need.

15 – Define the term "Relevant"?

If it is one that the user perceives as containing information of value with respect to their personal information need.

16 – What's the Information Retrieval System Effectiveness?

The quality of its search results.

17 – Mention two term which can measure the IR System effectiveness?

1 – Precision: What fraction of the returned results are relevant to the information need?

2 – Recall: What fraction of the relevant documents in the collection were returned by the system?

18 – What's the Grepping?

Process in UNIX used to retrieve desired information through text.

19– Why the Grepping is very effective process?

1. **Given the speed of modern computers.**
2. **Allows useful possibilities for wildcard pattern matching through the use of regular expressions.**

20 – Mention 3 shortfalls of Grepping?

1. To process large document collections quickly. The amount of online data has grown at least as quickly as the speed of computers.
2. To allow more flexible matching operations.
3. To allow ranked retrieval.

21 – Define Boolean Retrieval Model?

The Boolean Retrieval Model is a model for Information Retrieval in which we can pose a query which is in the form of a Boolean expression of terms (AND – OR - NOT).

22 – Problems with Boolean Retrieval Model?

1- Can't build the Matrix when we have a big collection of documents.
2- User don't like to write a Boolean Expression.
3- The Boolean model doesn't consider term weights in queries.
4- The result set of a Boolean query is often either too small or too big.

23 – What's the difference between Inverted Retrieval and Forward Retrieval?

- Forward Retrieval: Is the list of documents, and which words appear in them.
- Inverted Retrieval: Is the list of words, and the documents in which they appear.

24 – What's the Initial Stages of text processing?

- Tokenization:
- Normalization:
- Stemming:
- Stop Words:

25 – What's the Indexer Steps?

1. **Token Sequence.**
2. **Sort.**
3. **Dictionary and Postings.**

26 – How measure we the efficiency of IR system Implementation?

- **How do we index efficiently?**
- **How much storage do we need?**

27 – What's the goal of Extended Retrieval Model?

The goal of the Extended Boolean Model is to overcome the drawbacks of the Boolean Model that has been used in Information Retrieval.

28 – Define the Bi-Words Index with example?

Index every consecutive pair of terms in the text as a phrase.

- **Example: Stanford University Palo Alto can be broken into the Boolean.**
  Query on bi-words:
  - **Stanford University AND University Palo AND Palo Alto.**

29 – What's the issues of bi-word Index?

1. **False positives.**
2. **Index blowup due to bigger dictionary.**
3. **Bi-word indexes are not the standard solution (for all bi-words) but can be part of compound strategy.**

30 – Define the Positional Index with example?

- In the postings, store, for each term the position(s) in which tokens of it appear:

  <term, number of docs containing
  term;
  doc1: pos1, pos2,.....;
  doc2: pos1, pos2,.....;
  etc.>

Example

<be: 993353;
- 1: 7, 18, 33, 86, 231;
- 2: 3, 184;
- 4: 17,121, 303, 486, 531;
- 5: 363, 386, ...>

31 – Which algorithm is used in Phrase queries?

For phrase queries, we use a merge algorithm recursively at the document level.

32 - Rules of thumb?

1. A positional index is 2-4 as large as a non-positional index
2. A positional index size 35-50% of volume of original text (>) Positional index is about the size of 10% of the original text.
3. These rules of thumb will hold for English like language. Different languages may have different results.

33 – What was Williams et al at (2004) doing?

- Williams et al (2004) evaluated a more sophisticated mixed indexing scheme (Bi – Words + Positional):
  - A typical web query mixture was executed in ¼ of the time of using just a positional index
  - It required 26% more space than having a positional index alone.

34 – What's the problems in Jaccard Coefficient for Scoring?

1. It does not consider term frequency (How many times a term occurs in a document).
2. We need or sophisticate way of normalizing length.

35 – What's the problems Vector Representation, then define the Bag of Words Model?

- Vector representation does not consider the ordering of words.
- Bag of Words Model: Two Queries have the same vectors.

36 – Define the term Frequency $tf_{t,d}$?

The term frequency $tf_{t,d}$ of term t in the document d is defined as the number of times that t occurs in d.

37 – Why do we use $\log_{10}(N/df_t)$ instead of N/df_t in idf weight?

We use $\log_{10}(N/df_t)$ instead of N/df_t to "dampen" the effect of idf.

38 – Define the Collection Frequency?

The collection frequency of t is the number of occurrences of t in the collection, counting multiple occurrences.

39 – Define tf-idf weighting?

The tf-idf weight of a term is the product of its tf weight and its idf weight.

40 – What's the First Cut?

First Cut is Distance between the end points of the two vectors.

41 – Why Euclidean distance is a bad idea?

Because Euclidean distance is large for vectors of different lengths.

42 – Write the equation to calculate F-Measure?

F-Measure = (2 * Precision * Recall) / (Precision + Recall)