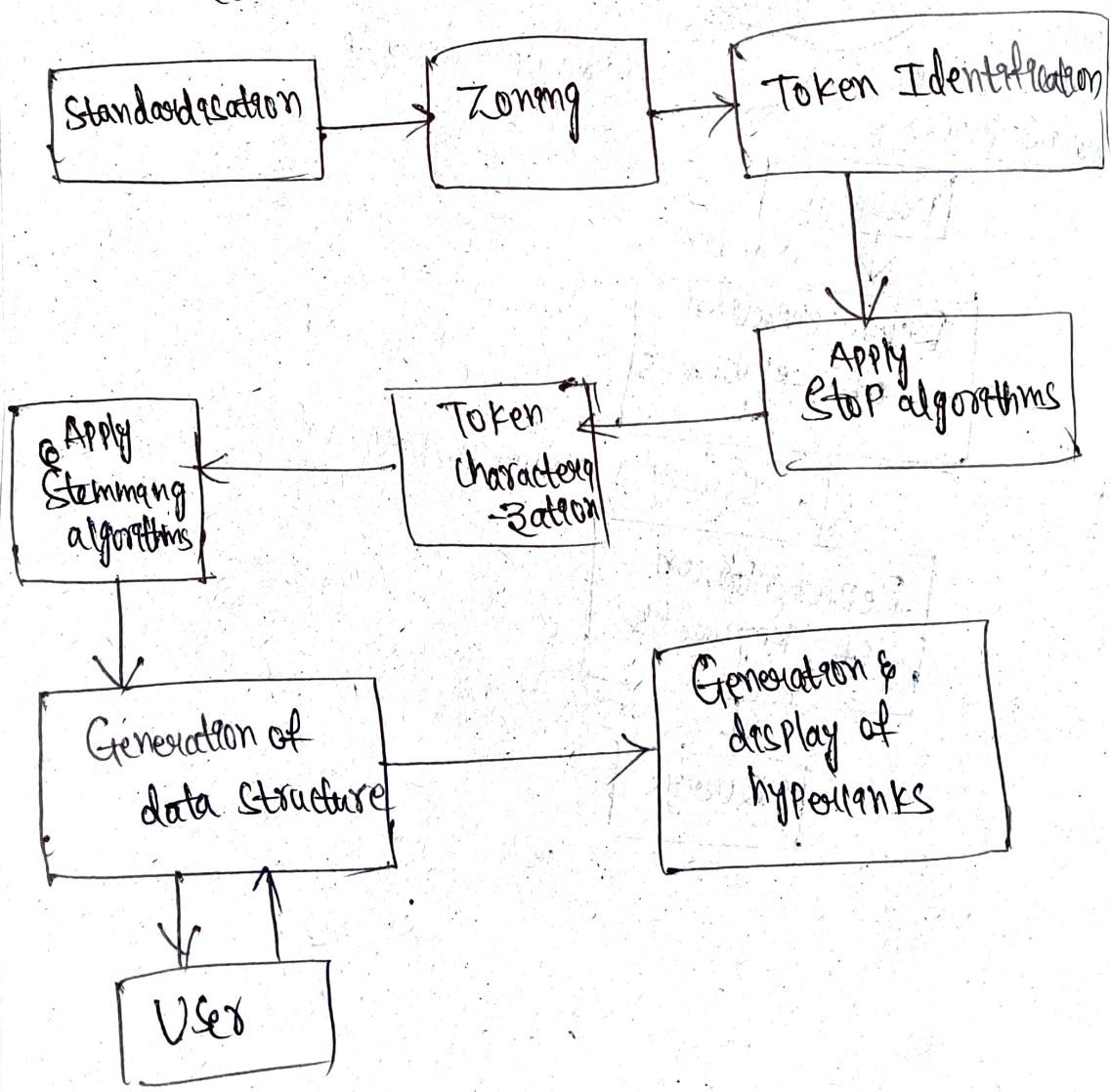
Unit-3

* Automatic Indexing

→ It can defined as an Indexing method.
→ It is one of the indexing methods.
→ It is used to represent a document.
→ Based on the index term, a document is represented.
→ Index terms include any of the following:
- Single words
- Lengthy phrases
- Combination of single words & lengthy phrases.



→ This process includes different steps:

1. Standardisation
2. Zoning
3. Token Identification
4. Apply Stop algorithms
5. Token Characterization
6. Apply Stemming algorithms

* Classes of Automatic Indexing

→ the classes of automatic indexing are:

- Statistical Indexing
- Natural Language Processing
- Concept Indexing
- Hypertext Indexing.

* Statistical Indexing

→ An Indexing technique that determines the significance of a document
  - by using a number called as "Statistical Indexing"

→ It is a class of automatic Indexing.

→ It uses number in this indexing.

→ there are two methods of Statistical indexing:
  1. Probabilistic Statistical Indexing
  2. Vector Statistical Indexing.

1. Probabilistic Statistical Indexing

→ Probabilistic Statistical Indexing is commonly referred as PSI.

→ It is also known as Probabilistic Weighting approach.

→ In this method, Significance of a document is estimated on the basis of probability value.

→ In other words, documents are ~~based~~ ranked based on the probability value.

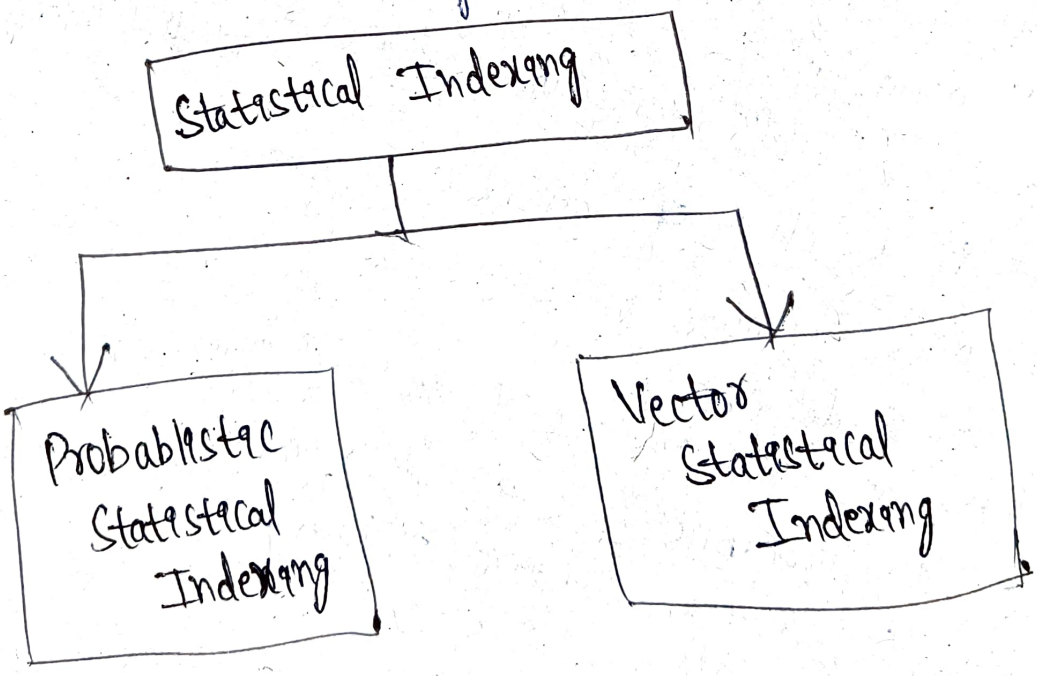
2. Vector Statistical Indexing

→ It is commonly referred as VSI

→ Significance of a document is estimated by using a vector.

→ It is also known as Vector Weighting approach

→ It includes :- Binary technique
- Weighted technique.



f any cluster. T_5 is similar to T_6 and T_8, so they all a... erm-relationship matrix for the ...

* Natural Language Processing

→ It is commonly referred as NLP

→ Natural language is nothing but User has giving emotions as well as user Expressing his views.

Ex: "Happy or Sad" "Good or Bad"

→ the goal of NLP is to use Semantic Information and Statistical Information.

→ NLP is a method, that is used to improve document indexing :
- Semantically
- Statistically.

→ thus improves the precision of searches.

Ex: 1. "Object-oriented Programming languages"

* By using Statistical approach, we get
- "Object-oriented Programming"
- "Programming languages"

* By using NLP, we get
- "Object-oriented language"
- "programming language"
- "Object-oriented Programming language".

2. "Information Retrieval System"

* Statistical approach
- "Information Retrieval"
- "Retrieval System"

* NLP
- "Information System"
- "Retrieval System"
- "Information Retrieval System"

* Concept Indexing

→ It is one of the class of automatic indexing.

→ Concept Indexing uses the words.

→ In NLP, Concepts in a document can be represented in following manners:
- Semantically
- Statistically.

→ Whereas in Concept Indexing, Concepts are represented in a more orthodox format.

→ Consider an example of a document which consists of a word "Computer",

→ It is related to different concepts like:
- Electronic device
- Storage device
- Processing device
- Programming

Word	Related concepts	Weight value (Assumed)
Computer	Electronic device	0.50
	Storage device	0.35
	Processing device	0.14
	Programming	0.22

* Hypertext Linkage Indexing

→ Hypertext linkage is one of the automatic indexing.

→ Hypertext is a data structure.

→ It is one of the most commonly used methods in IRS.

→ It is a special class of Indexing.

→ It helps to find relevant information.

→ It not only helps to find relevant information, but also additional information about it.

- When a user opens any search engine & type search statement
- Based on that statement, various sites will be displayed.
- these are nothing but hypertext documents.
- when user click on these links.
- You will enter the home page of that site.

* Clustering

→ Clustering means Grouping of similar objects that belong to the same subject.

→ Clustering improves the efficiency & effectiveness of retrieval process.

→ It groups similar objects into a class.

→ Clustering in IRS is of two types:

1. Term clustering
2. Document clustering
3. Item clustering.

1. Term clustering

→ this clustering done on similar terms.

→ this clustering is done to create statistical thesaurus.

→ Term clustering can be done manually or automatically

→ If Term clustering done automatically, it is knowns as Automatic Term clustering.

→ A Term may be word or group of words or a Single Paragraph

→ It increases recall.

2. Document clustering

→ It is one type of clustering.

→ It clusters the documents.

→ It is used create document clusters.

→ It is done on textual documents.

→ Also known as text clustering.

Ex: Cluster 1: Sports

→ Sports cluster consists of :

Doc 1 – football

Doc 2 – Cricket

Doc 3 – Basket Ball

Cluster 2: Politics

Doc 1 – Election updates

Doc 2 – Parties

* Item Clustering

→ It is similar to term clustering.

→ It is used to create statistical thesaurus.

→ Item clustering can be done either manually or automatically.

→ Item may be phrase, word, collection of words, diagram or a picture.

→ Similarity b/w documents is based on two items that have terms in common.

→ The similarity function is performed b/w rows of item matrix.

→ Default threshold value is 10.

$$S(\text{Item } i, \text{Item } j) = \Sigma (\text{Term } i, k)(\text{Term } j, k)$$

Ex: 10 is greater than are equal.

Item Id	Item1	Item2	Item3	Item4	Item5
1		11	3	6	22
2	11		12	10	36
3	3	12		6	9
4	6	10	6		11
5	22	36	9	11	

* Greater than or equal to 10, is an indicated as 1
* Less than 10, indicated as - 0

Id	Item1	Item2	Item3	Item4	Item5
1		1	0	0	1
2	1		1	1	1
3	0	1		0	0
4	0	1	0		1
5	1	1	0	1	

IRS → Results

IRS

* Hierarchy of Clusters

→ Hierarchy of cluster is a cluster technique.
→ It groups similar objects / terms into a set of clusters.
→ Here, each cluster is different from each other.
→ But objects in clusters are similar.
→ This technique groups the objects in the form of a tree.
→ the clusters are arranged in hierarchy manner.
→ Types are :

1. Hierarchical Agglomerative Clusters (HAC)

→ It begins with un-clustered item.
→ And perform pair-wise similarity measures.

2. Hierarchical divisive clustering

→ It starts with large cluster.
→ It breaks into smaller cluster.

→ It reduce overhead of search.
→ It perform top-down search
→ It provide visual representation

Visualizing the Hierarchy of Clusters Using Dendrograms

Graphical representation: