Page No.

Date : / /

Name : Dhiraj Bodake

Roll No : 18141216

TUTORIAL NO : 03

Q.1 What is cluster hypothesis? Explain it with R-R & R-N-R association.

→ Cluster Hypothesis:

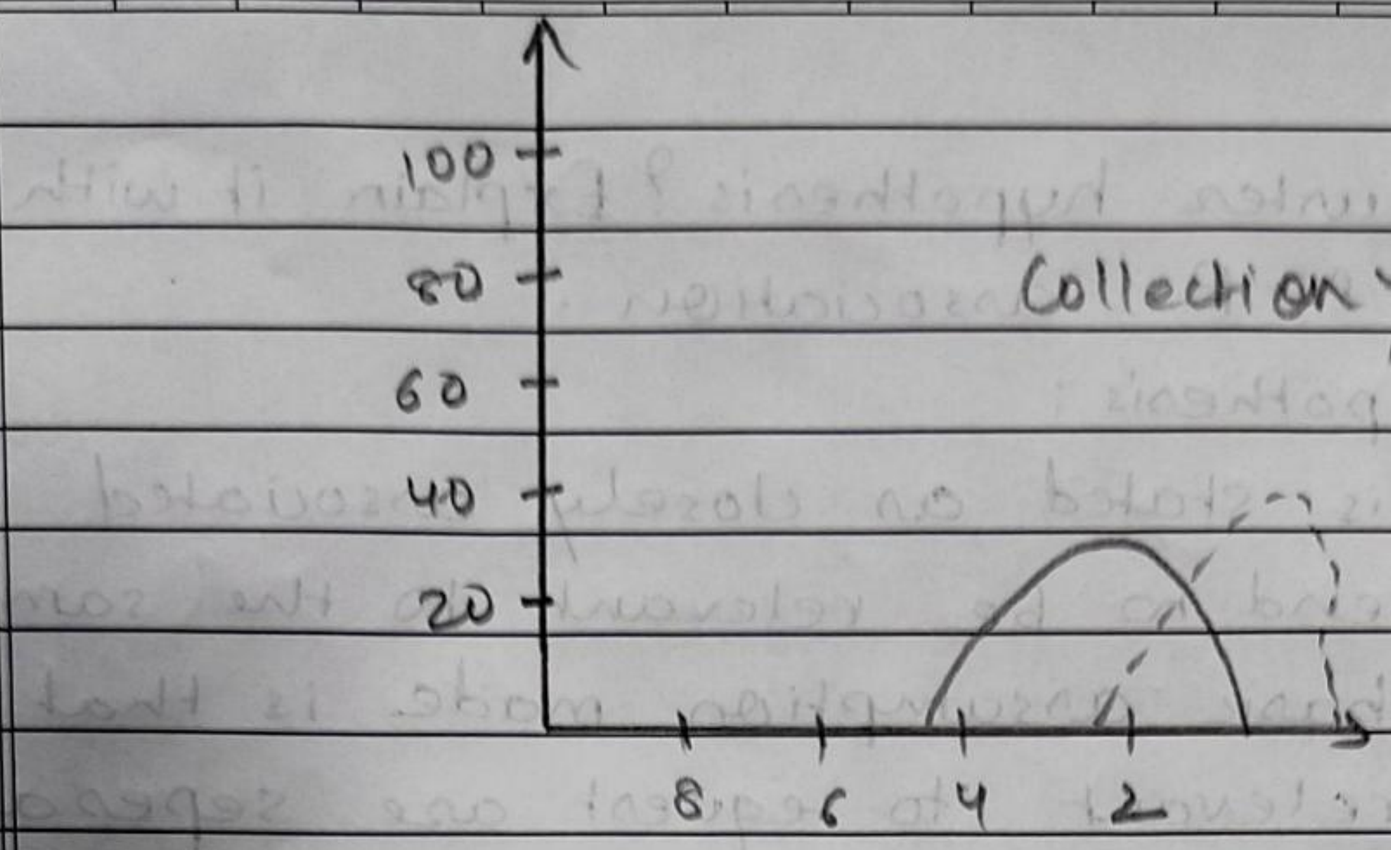
It is stated as closely associated documents tend to be relevant to the some request. A basic assumption made is that document relevant to request are seperated from those that are irrelevant.

This case be proved by computing the association between all pairs of documents.

a) Both of which documents are relevant to request

b) one of which is relevant & other are non-relevant.

- Plotting relative frequency against strength of association for two collection x & y.



- It is the seperation bet^n distribution that one attempts to exploit in document clustering.
- Increase distance between 2 distributions RR & RNR as we want to make it more likely for info. retreival of relevants docs.

Q.2] Explain inverted file as indexing technique.

→ Inverted file is word-oriented mechanism for indexing a text collection in order to speed up the searching task. Inverted file structure is composed of two elements : vocabulary & occurences.

- The vocabulary is the set of all different words in the text. For each such word a list of all the text positions where the word appears is stored. The set of all those lists is called the "occurences'. These positions can refer to words or characters

Page No.
Date

Word positions (position i refers to i^{th} word) simplify phrase & proximity queries while character positions. (position i is i^{th} character) facilitate direct access to matching text positions.

- Occurrences demands more space. Since each word appearing in the text is referred once in that structure, extra space is $O(n)$.
- To reduce space requirements, a technique called block addressing is used. Text divided in blocks & the occurrences point to the blocks where the word appears. The classical indices which point to the exact occurrences are called "full inverted indices".
- By using block addressing not only the pointers be smallest because there are fewer blocks than positions, but also all occurrences of word inside single block are collapsed to one reference.

Q.3. Build an inverted index for the following text. What is meant by vocabulary & occurrences? Explain it for given sample text.

'This is a text. A text has many words. Words are made from letters'

→ Given Text is written as:

1	6	9	11	17	19	24	28	33	40
This	is	a	text.	A	text	has	many	words.	Words
46	50	55	60						
are	made	from	letters.						

Inverted Text is :

Vocabulary	Occurrences
letters	60...
made	50...
many	28...
text	11, 19...
words	33, 40...

This is the inverted text built from sample text : Words are converted to lowercase; The occurrences pt. to character positions in text.

- Occurences : for each word, list of all text positions : where the word appears is stored. The set of all those lists is called occurrences.
- Vocabulary : Vocabulary is the set of different
- words in the text.

For given text the occurrences point to character positions in the text : Vocabulary are the words in the text i.e. 'letters', 'made', 'many', 'text', etc.