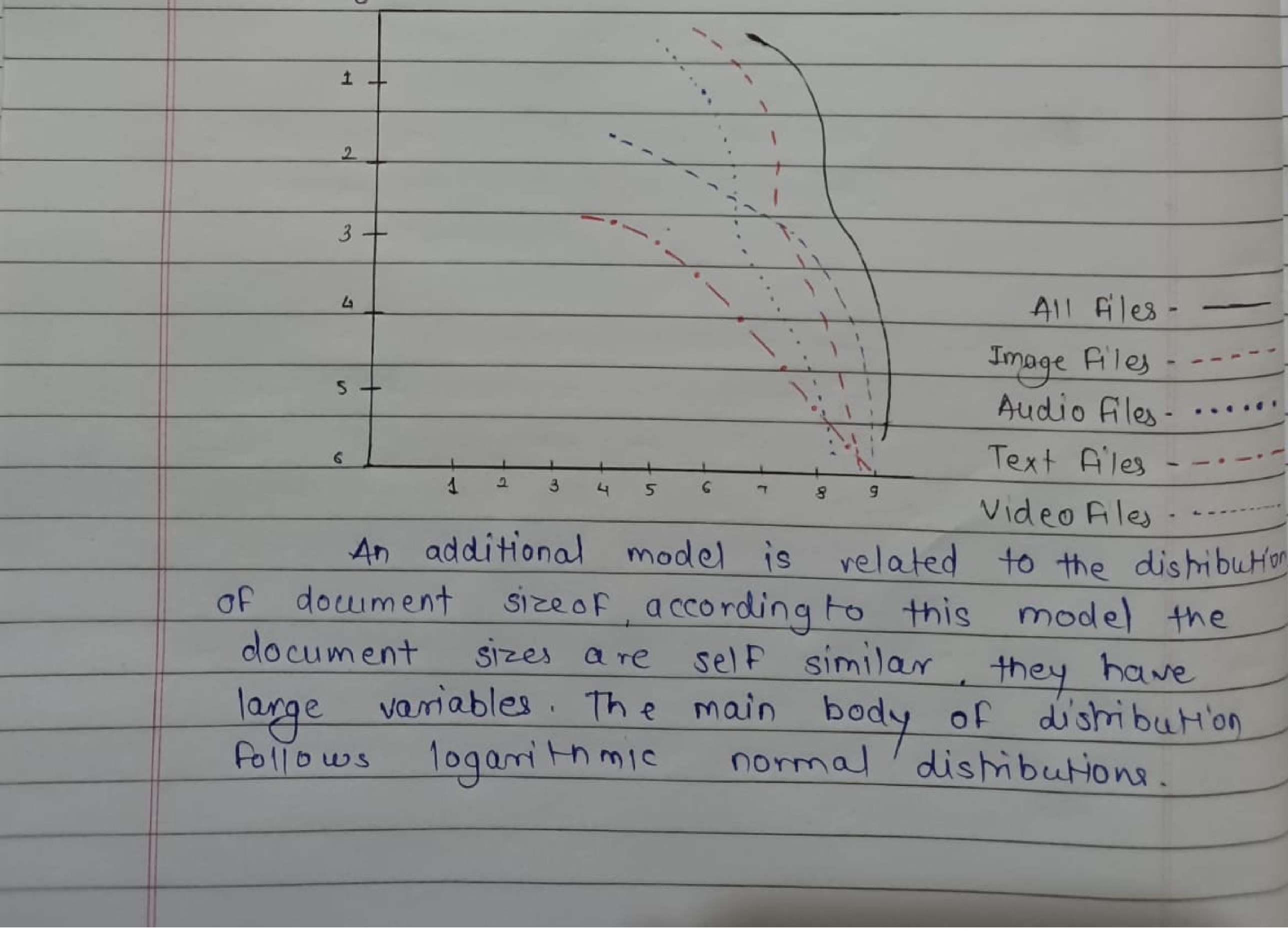
Name: Dhiraj Bodake

Roll No: 18141216

TUTORIAL NO : 11

Q.1 With the help of neet graph, explain modelling the web in details.

→ In particular, the vocabulary grows faster and word distributing should be more biased, however there are no experiments on large web collections to measure these parameters.



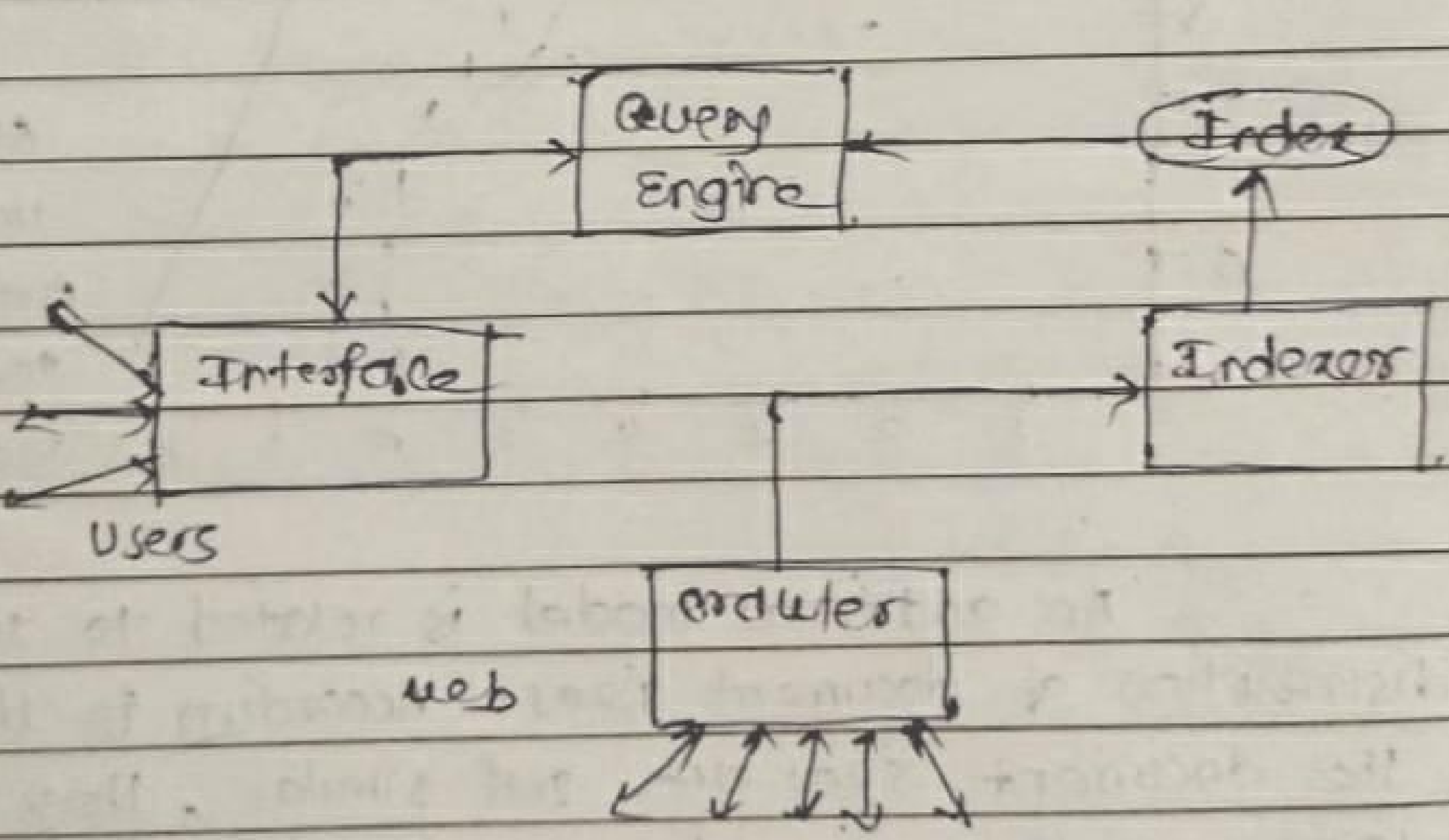
An additional model is related to the distribution of document sizeof, according to this model the document sizes are self similar, they have large variables. The main body of distribution follows logarithmic normal distributions.

Page No.:

Q.2 Explain typical crawler - indexes architecture with neat diagram

→ Crawlers are programs that traverse the web sending new or updated pages to a main server where they are indexed. Crawlers are also called robots, spiders, wanderers, walkers and knowbots

- Inspite of their name, a crawler does not actually move to and run on local system and send requests to remote web servers
- the index is used in a centralized basis to answer queries submitted from different places in the web. fig. below shows software arch of search engine based on Altavista architecture



Typical crawler - indexer architecture

- Architecture has two parts: one is that deals with users consisting of user interface and query engine and another that consist of crawler and indexer

Page No.:

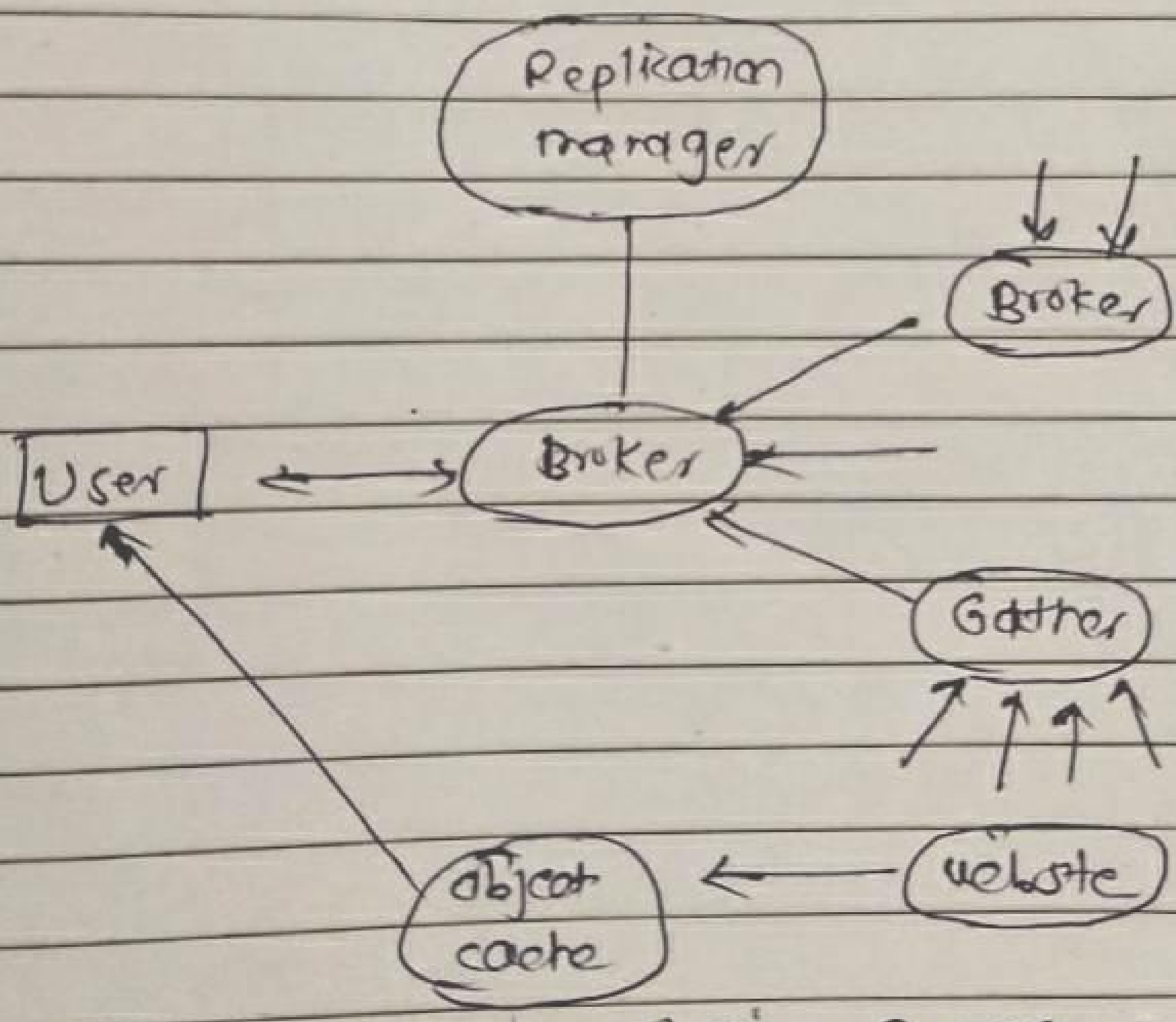
a.3 Explain in detail harvest as the distributed architecture with help of neat diagram

→ The harvest distributed approach addressed several of the problems of crawler - indexed architecture

① web servers receive requests from diff. crawlers increasing their load.

② Web traffic increases because crawlers retrieve entire objects, but most of their content is discarded.

③ Information is gathered independently by each crawler without coordination both all search engines.



- This arch allows sharing of work and info in very flexible and generic manner.
- One of the goals of harvest architecture is to build topic-specific broker