Internal Assessment Test 3 scheme and solution– Jan 2024

| Sub: | Big Data Analytics | | | | | Sub Code: | 18CS72 | Branch: | CSE |
|---|---|---|---|---|---|---|---|---|---|
| Date: | 4/1/2024 | Duration: | 90 mins | Max Marks: | 50 | Sem / Sec: | VII/A,B,C | | OBE |

Answer any 5 FIVE FULL Questions

| | | MARKS | CO | RBT |
|---|---|---|---|---|
| 1 | Explain Hive data models and Hive Integration and Workflow Steps with diagram.<br>Hive data model – 3 marks<br>Diagram -1 marks<br>Workflow steps- 6 marks | [10] | CO4 | L2 |
| 2 | How to find TF and IDF? (3 marks) Calculate TF-IDF using given Data Term t appears 55 times in a document. Document contains a total of 1000 words.<br><br>Collection of related documents contains 10,078 documents. 190 documents out of 10,000 documents contain the term t web content mining tasks (7 marks) | [10] | CO5 | L3 |
| 3 | Explain five phases in a process pipeline in Text mining.<br><br>Diagram- 1 marks<br><br>Pipeline phases-9 marks | [10] | CO5 | L2 |
| 4 | Explain the Euclidean, Jaccard and cosine distance.<br><br>Euclidean, Jaccard and cosine distance – 3+4+3 marks | [10] | CO4 | L4 |
| 5 | Fill in the blanks<br><br>a) A ---------- enables the link between two groups. (1 mark)<br>b) Detecting the ---------- leads to direct discovery of communities. (1 mark)<br>c) ---------- is the metric for measuring similarity between vertices of the same type. (1 mark)<br>d) ---------- is a way of defining the centrality of a vertex in reference to other vertices. (1 mark)<br>e) A ---------- is an index page that out-links to a number of content pages. An -------- is a page that has recognition due to its useful, reliable and significant information. ( 2 marks)<br>f) Explain association rule mining. (4 marks) | [10] | CO5 | L2 |
| 6 | Write difference between: Pig Vs Hive Vs MapReduce Vs SQL.<br>(2.5 marks each) | [10] | CO4 | L2 |

Solutions:

1 a) Hive data models

| Name | Description |
|---|---|
| Database | Namespace for tables |
| Tables | Similar to tables in RDBMS<br>Support filter, projection, join and union operations.<br>The table data stores in a directory in HDFS |
| Partitions | Table can have one or more partition keys that tell how the data stores |
| Buckets | Data in each partition further divides into buckets based on hash of a column in the table.<br>Stored as a file in the partition directory. |



| Step No. | OPERATION |
|---|---|
| 1 | Execute Query: Hive interface (CLI or Web Interface) sends a query to Database Driver to execute the query. |
| 2 | Get Plan: Driver sends the query to query compiler that parses the query to check the syntax and query plan or the requirement of the query. |
| 3 | Get Metadata: Compiler sends metadata request to Metastore (of any database, such as MySQL). |
| 4 | Send Metadata: Metastore sends metadata as a response to compiler. |
| 5 | Send Plan: Compiler checks the requirement and resends the plan to driver. The parsing and compiling of the query is complete at this place. |
| 6 | Execute Plan: Driver sends the execute plan to execution engine. |
| 7 | Execute Job: Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Then, the query executes the job. |
| 8 | Metadata Operations: Meanwhile the execution engine can execute the metadata operations with Metastore. |
| 9 | Fetch Result: Execution engine receives the results from Data nodes. |
| 10 | Send Results: Execution engine sends the result to Driver. |
| 11 | Send Results: Driver sends the results to Hive Interfaces. |



Components of Hive architecture

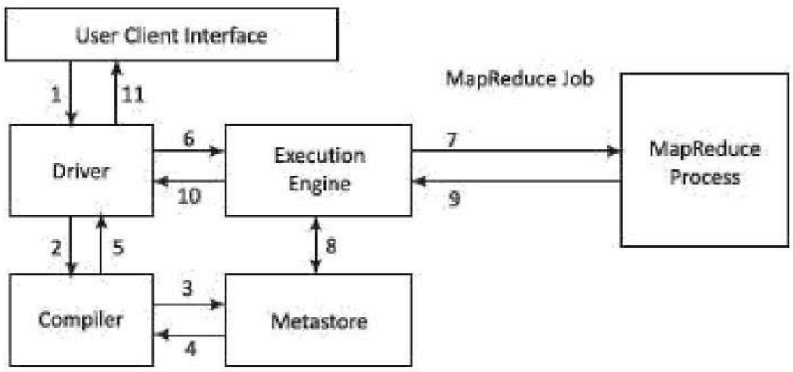



**Figure 4.11** Dataflow sequences and workflow steps

are:

Q2.

Tf/Idf

$$\text{TF-IDF}(t) = \frac{\text{No. of times t appears in a document}}{\text{Total No. of terms in the document}} \times \log \frac{\text{No. of documents in the collection}}{\text{No. of documents that contain t}}. \quad (9.1)$$

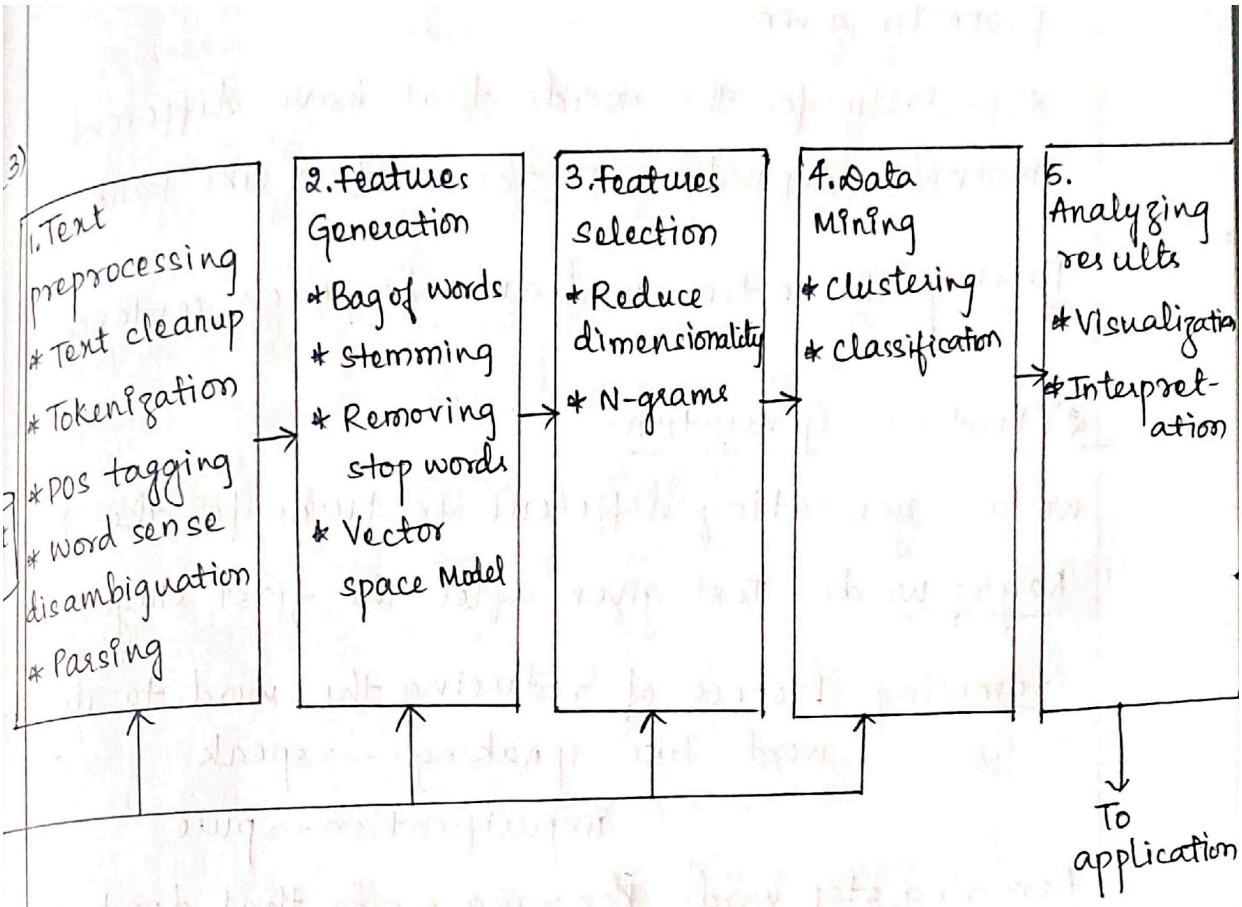
(i) TF-IDF belongs to Vector space Model.

It is a method used to find the importance of a word in the text.

TF-IDf is a numerical measure.

If a word is repeated many times in the document, then it is an important word in the document whose TF-IDF score will be high.

The words that infrequently occur in the document, their importance is very low so TF-IDF score is low as well.




## 1) Text preprocessing

Raw text is given as input to this stage.
first text cleanup is done where unnecessary/ unwanted data is removed from the text.

Tokenization - text is divided into tokens (words)

POS tagging - Part of speech is assigned to each token in the text to be able to identify names of people, places etc.
It could be Noun, Verb, Adverb, Adjective, etc.

$$\text{TF-IDF} = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Numbers of terms in the document}} \times \log_{10} \frac{\text{Number of documents in the collection}}{\text{Number of documents that contain the term } t}$$

Above is the formula to calculate TF-IDF

Given:

Number of times term t appears in a document = 20

Number of terms in the document = 100

Number of documents in the collection = 10,000

Number of documents that contain term t = 100

$$\therefore \text{TF-IDF} = \frac{20}{100} \times \log_{10} \frac{10,000}{100}$$

$$= 0.2 \times \log_{10}(100)$$

$$= 0.2 \times 2$$

$$\boxed{\text{TF-IDF} = 0.4}$$

Word sense Disambiguation - Meaning of each word is properly given.
Especially for the words that have different meanings depending on the context like bank, etc.

Parsing - Parse tree is drawn for each sentence.

2) Features Generation

We are generating different features for the text.

Bag of words - Text given after the first stage.

Stemming - Process of reducing the word to its root word like speaking → speak
Impurification → pure

Removing stop words - Removing words that don't matter like for, at, from

Vector Space Model - tf-Idf model to find Importance of a word/term.

3) Features Selection

Selecting a few important features out of the many that are generated.

* Reduce dimensionality in the text.
* N-grams - finding words of length N.
2-gram can be ['Beautiful Flower', 'That Rose']

4) Data Mining

clustering - unsupervised method to cluster similar words in the text.

classification - supervised, to classify the words.

5) Analyzing Results

In this step, the results are analyzed

visualization - Graphs and other type of visualization is done to analyze data.

Interpretation - of the data

Q4.

4) Jaccard distance

Let A and B be two different users who similarity we are trying to find

$$J = \frac{|A \cap B|}{|A \cup B|}$$

To find Jaccard similarity index, we use

$$J_{index} = 1 - \frac{|A \cap B|}{|A \cup B|} \times 100\%$$

$\cap$ - represents intersection
common things present in both A and B.

$\cup$ - represents Union
all things / words present in both A and B.

Eg: There are 100 Youths (Y) and 200 families (F).
40 out of 100 are interested to buy a car named 'XYZ'

$$J(Y, F) = \frac{40}{(100+200)} \times 100 = 13.33\%$$

## Euclidean distance

$$D_{Eu} = \left[ \sum_{i=1}^{v} (x_i - x_i')^2 \right]^{1/2}$$

This can also be used to find the distance and then use the distance to find similarities between the users.

## Cosine distance

Let v and V be two vectors

$$D_{cos} = \frac{\sum_i v_i V_i}{\sqrt{\sum_i v_i^2} \sqrt{\sum_i V_i^2}}$$

$D_{cos}$ formula can be used for the purpose of finding similarities as well.

Q5 a) bridge

b) clique

c) simrank

d) Graph vertex closeness Cc (v

e) hub and authority

**Association rule** The rules enable relating the pages, which are most often referenced together in a single server session. These pages may not be directly connected to one another using the hyperlinks

Other uses of association rule mining are:

(i) Reveal a correlation between users who visited a page containing similar information.
(ii) Provide recommendations to purchase other products. For example, recommend to user who visited a web page related to a book on data analytics, the books on ML and Big Data analytics also.
(iii) Provide help to web designers to restructure their websites.
(iv) Retrieve the documents in prior in order to reduce the access time when loading a page from a remote site.

Q6.

| Pig | SQL |
|---|---|
| Pig Latin is a procedural language | A declarative language |
| Schema is optional, stores data without assigning a schema | Schema is mandatory |
| Nested relational data model | Flat relational data model |
| Provides limited opportunity for Query optimization | More opportunity for query optimization |

**Table 4.15** Differences between Pig and Hive

| Pig | Hive |
|---|---|
| Originally created at Yahoo | Originally created at Facebook |
| Exploits Pig Latin language | Exploits HiveQL |
| Pig Latin is a dataflow language | HiveQL is a query processing language |
| Pig Latin is a procedural language and it fits in pipeline paradigm | HiveQL is a declarative language |
| Handles structured, unstructured and semi-structured data | Mostly used for structured data |

| Pig | MapReduce |
| --- | --- |
| A dataflow language | A data processing paradigm |
| High level language and flexible | Low level language and rigid |
| Performing Join, filter, sorting or ordering operations are quite simple | Relatively difficult to perform Join, filter, sorting or ordering operations between datasets |
| Programmer with a basic knowledge of SQL can work conveniently | Complex Java implementations require exposure to Java language |
| Uses multi-query approach, thereby reducing the length of the codes significantly | Require almost 20 times more the number of lines to perform the same task |
| No need for compilation for execution; operators convert internally into MapReduce jobs | Long compilation process for Jobs |
| Provides nested data types like tuples, bags and maps | No such data types |