What is the basic idea behind the **Vector Space Model**?

* Each query and each document are represented as vectors in a vector space (the term space). Documents are ranked according to decreasing similarity between document and query vectors (the most similar first).

Assume you want to create a Web search engine and someone provides you with all the necessary resources (machines, storage, etc.). Which problem would you face initially?

* Building the repository. We don't know how many Web pages are there and which they are.

What is one advantage of the **Vector Space Model** compared to the simple Boolean model we discussed in the lecture?

* It creates a ranked list of documents

Which one of the following statements about Information Retrieval is true?

* Information Retrieval aims at retrieving documents that are relevant with respect to a user's information need.

What are the 4 steps of indexing in **Information Retrieval**, as discussed in the lecture?

* Tokenization, Stopword elimination, Stemming, Creation of the inverted index

What are the steps required for data analysis?

* Select Technique, Build Model, Evaluate

Amazon will often show you what "Customers who viewed this item also bought".What type of analysis technique could this be based on?

* Association Analysis

Which of the following are examples of how to address data quality issues?

* Generate best estimates for invalid values
* Remove outliers
* Merge duplicate records
* Remove data with missing values

What is done to the data in the *Prepare* stage?

* Understand the data and preliminary analysis

What does the statistical function *range* do?

* Measures the difference between the largest and smallest value in a column

If today's temperature is 12oc What type of measurement is the temperature value?

* Interval

The following is an example of the Likert Scale:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Like** | **Almost Like** | **Neutral** | **Almost Dislike** | **Dislike** |
| 1 | 2 | 3 | 4 | 5 |

A survey could ask "*What do you think of Big Data?*" with the top row representing the possible responses. The results from the survey can be coded as a number seen in the second row.

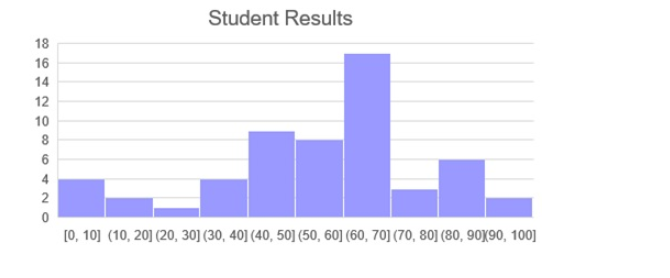
What type of data does this represent?

* Ordinal data

 What does the statistical function *Median*do?

* Finds the middle value in a data set

What type of graph is this an example of:



* Histogram

One approach to handling *dirty data* is to Fix it. What would this entail?

* Replace the incorrect value with the correct value.

Given a collection called *depts* that contain all the University’s departments and has the fields: *name, budget* and *location*. What is the MongoDB command to list the details of the department named “Maths and Computer Science”:

* db.depts.find({"name": "Maths and Computer Science"})

Given a collection called *depts* that contain all the University’s departments and has the fields: *name*, *budget* and *location*.What is the MongoDB command that will return all the departments with a budget greater than £10,000:

* db.depts.find({"budget": {$gt: 10000}})

Using the depts collection, what is the syntax in MongoDB to produce a sum of the budgets by location:

* db.depts.aggregate([

{ $group: { \_id: "$location", total: { $sum: "$budget" } } }

])

Given a collection called *myModules* that has the fields: number, name and credits, what is the command to add a new record for 6CS030:

* db.myModules.insert({

moduleno: "6CS030",

name: "Big Data",

credits: 20

})

MongoDB uses a concept called an *Aggregation Pipeline* to transform documents into aggregated results. Which SQL concept is this similar to?

* GROUP BY

You want to count how many documents there are in the *emp* collection. Which command should you use?

* db.emp.count()

You have a new collection called *myTweets* which has over 10,000 documents. You do not know what type of data it contains. Which command can help you find out the structure of a document?

* db.myTweets.findOne()

What command can be used to find documents containing the word *icy* in the *weather* collection. *icy* should be found no matter what case it is in (upper, lower, etc).

* db.weather.find({ text: { $in: [/sun/, /rain/, /icy/] }})

What MongoDB command is the equivalent of the SQL query:

SELECT deptno, avg(sal) AS avgSal  
FROM emp  
GROUP BY deptno

* db.emp.aggregate ( [  
  { $group:  
  { \_id: "$deptno", avgSal: {$avg: "$sal"} } }  
  ])

In Spark transformations are *lazily evaluated*. What does this mean?

* The transformation is not executed until an action needs the result

The following is a snipped of Java code from the *main* method for the Word Count program:

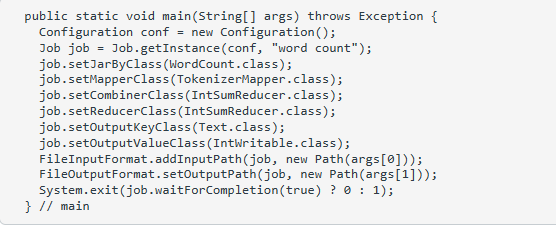
A screen shot of a computer code

Description automatically generated

Which line of code tells Hadoop which Mapper class to use?

* job.setMapperClass(TokenizerMapper.class);

The following is a snipped of Java code from the *main* method for the Word Count program:



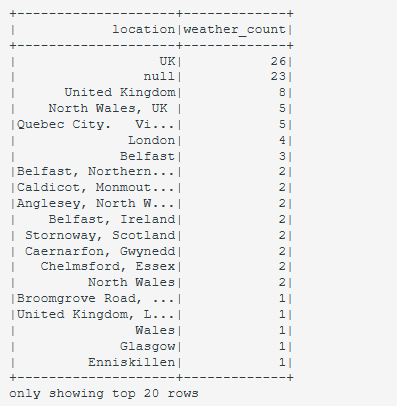
Which line of code tells Hadoop what output directory to create?

* FileOutputFormat.setOutputPath(job, new Path(args[1]));

Assuming you have a directory called*input\_dir* created in the Hadoop dfs. Which command allows you to view what files it contains?

* hdfs dfs -ls input\_dir

The following is some example output generated in Spark using the Weather dataset:



Which of the following commands produced the output above?

* spark.sql("SELECT user.location, count(\*) as weather\_count FROM weather GROUP BY user.location ORDER BY weather\_count desc").show()

Using a Spark DataFrame and the Weather json file, which of the following commands would show just the screen name and name of tweets made by German users:

* df.filter(df['user.lang'] == "de").select('user.screen\_name', 'user.name').show(40)

What is the order of the three steps to Map Reduce?

* Map -> Shuffle and Sort -> Reduce

Which of the following Apache Projects can also be viewed as a NoSQL database:

* HBase

Which of the following can be used to provide Machine Learning in Apache Spark:

* MLib

Which of the following can be used to analyse a continuous stream of data in Apache Spark:

* Spark Streaming

What do you call a computer network where data is stored on more than one node, which may be replicated

* distributed file system

What sort of data is best suited to a bitmap index:

* Low-cardinality columns such as gender

The data in a Data Warehouse is rarely deleted, because the data represents the company’s history. What stage of Bill Inmon’s famous quote is this an example of:

* Non-volatile

Which of the following types of applications are most suitable for an Online Transactional Processing (OLTP) system:

* Student Information System such as eVision

Is the original data modified during the HDFS lifecycle?

* No

What is the fundamental unit of data in Apache Spark?

* RDD (Resilient Distributed Dataset)

HDFS stands for Hadoop Distributed File System,

RSS stands for Really Simple Syndication

What are two types of operations in Apache Spark?

* Actions and Transformations

What is a Data Lake?

* A physical instantiation of a logical Data Warehouse

Which of the following types of applications are more suitable for Big Data technologies:

* Massive Grid Computer System such as CERN’s Large Hadron Collider Computing Grid

Which of the following is not a type of NoSQL database:

* Graphical database

Which architecture is best for a Big Data application:

* Distributed File System

What is a Commodity Cluster with respect to Big Data?

* A collection of computing nodes connected over a network

What animal is not related to any part of the basic Hadoop Stack '*Zoo*'?

* Horse

Name the high-level language that is a main part of Apache Pig

* Pig Latin

Which of the following is not a valid command to handle data in HDFS?

* cp -r /user/data /user/test/

What is the organising data structure for map/reduce programs?

* A list of identification keys and some value associated with that identifier

In the Word Count examples, in terms of key/values, what is the key?

* The word itself

In the Java version of the Word Count programs, which of the following examples of code would tell Hadoop which Mapper to use:

* Job job = Job.getInstance(conf, "word count");  
  job.setMapperClass(TokenizerMapper.class);

Which of the following requirements are needed for programming Big Data:

All of the answers

Handle fault tolerance

Distribute computations to nodes

Access data fast