Content

Quiz 1 - Introduction to Big Data

1. What are the main four dimensions associated with Big Data?
   Ans : Volume, Velocity, Variety, Veracity

2. What does Volume mean in the context of Big Data:
   Ans : Refers to the vast amount of data generated every second

3. What does Velocity mean in the context of Big Data:
   Ans : Refers to the speed at which new data is generated

4. What does Variety mean in the context of Big Data:
   Ans : Refers to the different types of data that can be used

5. What does Veracity mean in the context of Big Data:
   Ans : Refers to the trustworthiness of the data

6. According to Eric Schmidt in 2010, how much data is generated every two days:
   Ans : 5 Exabyte (EB)

7. Given 1 Kilobyte is 1000 bytes. How many multiples of kilobytes is an Exabyte (EB)
   Ans : 1000^6

8. Which of the following is an example of conversational data:
   Ans : Twitter feed

9. Which of the following is an example of sensor data:
   Ans : F1 telemetry information

10. Which of the following is an example of astronomical data:
    Ans : Radio waves collected from Jodrell Bank's Lovell telescope

11. Which of the following is an example of photo and video image data:
    Ans : Digital image

12. Which of the following types of applications are most suitable for a Big Data

system:
Ans : Massive Grid Computer System such as CERN's Large Hadron Collider Computing Grid

Quiz 2 - Big Data Process

1. What are the steps required for data analysis?
   Ans : Select Technique, Build Model, Evaluate

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

3. Which of the following are examples of how to address data quality issues?
   Ans : Remove outliers
   Ans : Remove data with missing values
   Ans : Merge duplicate records
   Ans : Generate best estimates for invalid values

4. What is done to the data in the Prepare stage?
   Ans : Understand the data and preliminary analysis

5. What does the statistical function range do?
   Ans : Measures the difference between the largest and smallest value in a column

6. If today's temperature is 12oc .What type of measurement is the temperature value?
   Ans : Interval

7. The following is an example of the Likert Scale:

Like	Almost Like	Neutral	Almost Dislike	Dislike
1	2	3	4	5

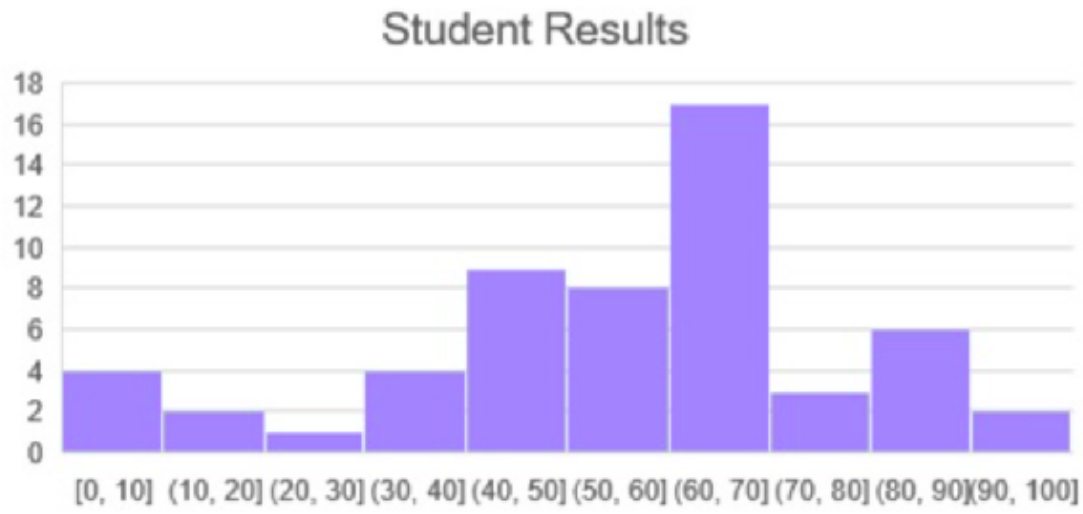
A survey could ask "What do you think of Big Data?" with the top row

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?
Ans : Ordinal data

8. What does the statistical function Median do?
   Ans : Finds the middle value in a data set

9. What type of graph is this an example of:



   Ans : Histogram

10. One approach to handling dirty data is to Fix it. What would this entail?
    Ans : Replace the incorrect value with the correct value

Quiz 3 - Types of Data

1. What type of data characteristics is shown in this example:

DEPT Table

DEPTNO	DNAME	LOC
10	ACCOUNTING	NEW YORK
20	RESEARCH	DALLAS
30	SALES	CHICAGO
40	OPERATIONS	BOSTON

Ans : Structured data

2. What type of data characteristics does the following suggest:

```
DEPT: {
    deptno: 10,
    dname: 'ACCOUNTING',
    loc: 'NEW YORK',
    auditor: 'KPMG'
    }
```

```
DEPT: {
    deptno: 30,
    dname: 'SALES',
    loc: 'CHICAGO',
    budget: 50,000
    }
```

```
DEPT: {
    deptno: 20,
    dname: RESEARCH',
    loc: 'DALLAS',
    prof: 'Prof Green',
    status: 'Silver'
    }
```

```
DEPT: {
    deptno: 40,
    dname: 'OPERATIONS',
    loc: 'BOSTON'
    }
```

Ans : Semi-structured data

3. What does the term NoSQL represent?
Ans : Not Only SQL

4. Which of the following is not an example of a type of NoSQL Database?
Ans : Relational database

5. Given the following data, what type of NoSQL database would suitable:

Key	Value
10	{Clark, King, Miller}
20	{Adams, Ford, Jones, Scott, Smith}
30	{Allen, Blake, James, Martin, Turner, Ward}
40	

Ans : Key-value database

6. Given the following data, what type of NoSQL database would suitable:NoSQL Structure.

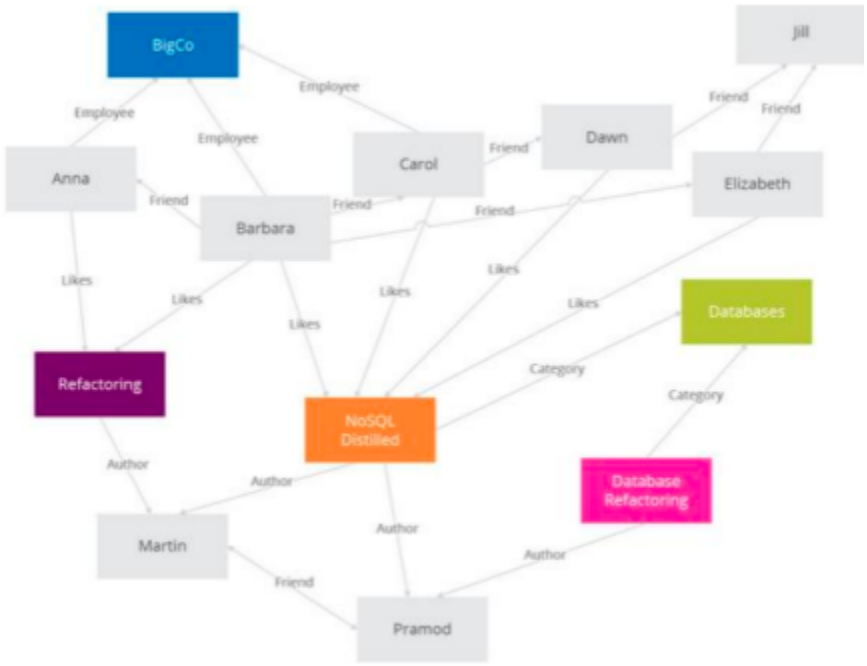
Ans : Document-based database

7. Given the following data, what type of NoSQL database would suitable:

EMPNO	ENAME	MGR	HIREDATE	SAL	DEPTNO	
7876	ADAMS	7788	19-Nov-15	1100	20	
EMPNO	ENAME	MGR	HIREDATE	SAL	COMM	DEPTNO
7499	ALLEN	7698	20-Feb-95	1600	300	30
EMPNO	ENAME	HIREDATE	SAL			
7839	KING	17-Nov-80	5000			
EMPNO	ENAME	MGR	HIREDATE	SAL	DEPTNO	EMAIL
7788	SCOTT	7566	16-Oct-15	3000	20	scott.tiger@oracle.com

Ans : Column-family database

8. Given the following data, what type of NoSQL database would suitable:



Ans : Graph-based database

9. Given a collection called myModules, what would be the MongoDB command to list all the data in this collection:

Ans : db.myModules.find()

10. What type of NoSQL data is MongoDB an example of?
    Ans : Document database

Quiz 4 - MongoDB

1. 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":
   Ans : db.depts.find({"name": "Maths and Computer Science"})

2. 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:
   Ans : db.depts.find({"budget": {$gt:10000}})

3. Using the depts collection, what is the syntax in MongoDB to produce a sum of the budgets by location:
   Ans :
   db.depts.aggregate ([
   { $group: { _id: "$location", total: {$sum: "$budget"} } }
   ])

4. Given a collection called myModules that has the fields: number, name and credits, what is the command to add a new record for 6CS030:
   Ans :
   db.myModules.insert({
   moduleno: "6CS030",
   name:"Big Data",
   credits:20}
)

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

Ans : GROUP BY

6. You want to count how many documents there are in the emp collection. Which command should you use?
   Ans : db.emp.count()

7. 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?
   Ans : db.myTweets.findOne()

8. 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).
   Ans : db.weather.find({text: /icy/i})

9. What command could be used to find several values in the text field of the weather collection:
   Ans : db.weather.find({ text: { $in: [/sun/, /rain/, /icy/] }})

10. What MongoDB command is the equivalent of the SQL query:
    SELECT deptno, avg(sal) AS avgSal
    FROM emp
    GROUP BY deptno
    Ans :
    db.emp.aggregate ([
    { $group:
    { _id: "$deptno", avgSal: {$avg: "$sal"} } }
    ])

Quiz 5 - Hadoop

1. Which architecture is best for a Big Data application:

Ans : Distributed file system

2. What is a Commodity Cluster with respect to Big Data?
   Ans : A collection of computing nodes connected over a network

3. What animal is not related to any part of the basic Hadoop Stack 'Zoo'?
   Ans : Horse

4. What does HDFS stand for?
   Ans : Hadoop Distributed File System

5. Name the high level language that is a main part of Apache Pig
   Ans : Pig Latin

6. Which of the following is not a valid command to handle data in HDFS?
   Ans : cp -r /user/data /user/test/

7. What is the organizing data structure for map/reduce programs?
   Ans : A list of identification keys and some value associated with that identifier

8. In the Word Count examples, in terms of key/values, what is the key?
   Ans : The word itself

9. Which of the following requirements are needed for programming Big Data:
   - ☐ Handle fault tolerance
   - ☐ Access data fast
   - ☐ Distribution computations to nodes
   - ☐ All of the answers

   Ans : All of the answers

Quiz 6 - Hadoop Zoo

1. In Spark transformations are lazily evaluated. What does this mean?
   Ans : The transformation is not executed until an action needs the result

2. The following is a snippet of Java code from the main method for the Word Count program:

```
public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
} // main
```

   Which line of code tells Hadoop which Mapper class to use?
   Ans : job.setMapperClass(TokenizerMapper.class);

3. The following is a snippet of Java code from the main method for the Word Count program:

```
public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
} // main
```

   Which line of code tells Hadoop what output directory to create?
   Ans : FileOutputFormat.setOutputPath(job, new Path(args[1]));

4. Assuming you have a directory called input_dir created in the Hadoop dfs. Which command allows you to view what files it contains?
   Ans : hdfs dfs -ls input_dir

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

```
+--------------------+-------------+
|            location|weather_count|
+--------------------+-------------+
|                  UK|           26|
|                null|           23|
|      United Kingdom|            8|
|    North Wales, UK |            5|
|Quebec City.   Vi...|            5|
|              London|            4|
|             Belfast|            3|
|Belfast, Northern...|            2|
|Caldicot, Monmout...|            2|
|Anglesey, North W...|            2|
|    Belfast, Ireland|            2|
| Stornoway, Scotland|            2|
| Caernarfon, Gwynedd|            2|
|   Chelmsford, Essex|            2|
|         North Wales|            2|
|Broomgrove Road, ...|            1|
|United Kingdom, L...|            1|
|               Wales|            1|
|             Glasgow|            1|
|         Enniskillen|            1|
+--------------------+-------------+
only showing top 20 rows
```

Which of the following commands produced the output above ?
Ans : spark.sql("SELECT user.location, count(*) as weather_count FROM weather GROUP BY user.location ORDER BY weather_count desc").show()

6. 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:
Ans : df.filter(df['user.lang'] == "de").select('user.screen_name', 'user.name').show(40)

7. What is the order of the three steps to Map Reduce?
Ans : Map -> Shuffle and Sort -> Reduce

8. Which of the following Apache Projects can also be viewed as a NoSQL database:
Ans : HBase

9. Which of the following can be used to provide Machine Learning in Apache

Spark
Ans : MLib

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

Quiz 7 - Consolidation

1. What do you call a computer network where data is stored on more than one node, which may be replicated?
Ans : distributed file system

2. What sort of data is best suited to a bitmap index:
Ans : Low-cardinality columns such as gender

3. What is the fundamental unit of data in Apache Spark?
Ans : RDD

4. What are two types of operations in Apache Spark?
Ans : Actions and Transformations

5. What is a Data Lake?
Ans : A physical instantiation of a logical Data Warehouse

6. Which of the following types of applications are more suitable for Big Data technologies?
Ans : Massive Grid Computer System such as CERN's Large Hadron Collider Computing Grid

7. Which of the following is not a type of NoSQL database?
Ans : Graphical database

8. 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 ?
Ans : Non-volatile

9. Which of the following types of applications are most suitable for an Online Transactional Processing (OLTP) system ?
   Ans : Student Information System such as eVision

10. Is the original data modified during the HDFS lifecycle?
    Ans : No

IR Test Quiz - Part 1

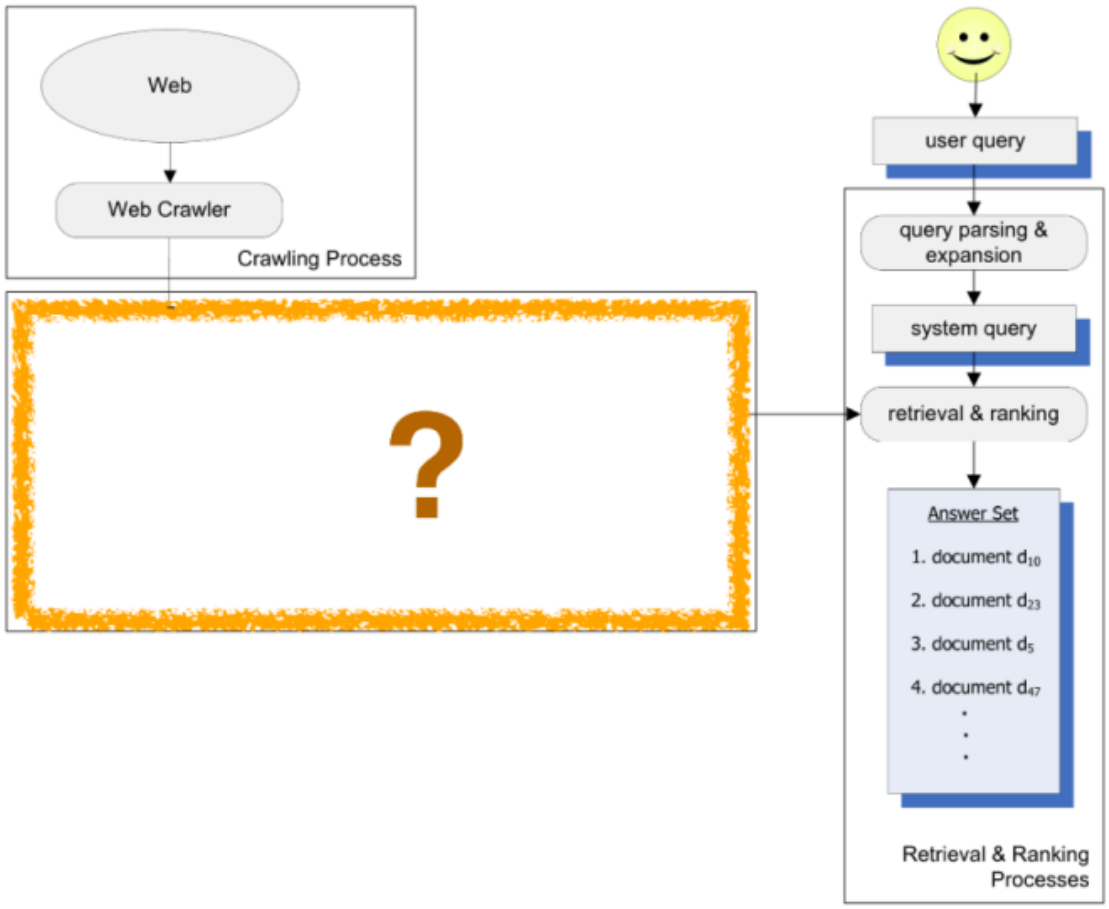
1. What is a Crawler (or Spider) for ?
   Ans : A crawler collects Web sites, interacts with Web servers and following links to new web pages to build and update a repository.

2. What are the 4 steps of indexing in IR, as discussed in the lecture ?
   Ans :
   Tokenization
   Stopword elimination
   Stemming
   Creation of the inverted index

3. Which one of the following statements about IR is true ?
   Ans: IR aims at retrieving documents that are relevant with respect to a user's information need.

4. How does a crawler (or spider) work ?
   Ans : Starting from a set of seed documents, the crawler follows looks and adds all found documents to the repository.

5. Below is the (incomplete) software architecture of a (Web) search engine as presented in the lecture. What is the name and purpose of the missing component (with the orange question mark) ?



Ans : Indexing Process - an index processes the document collection (or repository) and creates the inverted index.

6. What is the basic idea behind the vector space model ?
   Ans : 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).

7. Given the below inverted index, which documents would be retrieved for the query

```
retrieval systems OR multimedia OR images
```

retrieval systems

DOCID	POS
d1	10,25, 37
d2	23
d4	45,55
d7	3,56

multimedia

DOCID	POS
d4	18,20
d6	22,27
d7	5
d9	3,96, 127

images

DOCID	POS
d1	14,68
d3	29,73, 235
d4	11,145
d7	10

Ans : d1, d2, d3, d4, d6, d7, d9

8. What is the purpose of stopword elimination ?
   Ans : Frequent terms like "and", "or", don't bear any meaning for search and retrieval. This means we can eliminate them to keep our vocabulary small.

9. Consider the following query vector q and document vectors d1 and d2. According to the vector space model, what would be the scores of d1 and d2 and how would the documents be ranked ?

$$\vec{q} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

$$\vec{d_1} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \vec{d_2} = \begin{pmatrix} 1 \\ 0.5 \\ 1 \end{pmatrix}$$

Ans :
1. d2 with score 2
2. d1 with score 1

10. What is one advantage of the PageRank algorithm ?

Ans : It computes authority values of web pages offline as it doesn't depend on a query.

IR Test Quiz - Part 2

1. Given the below inverted index, which documents would be retrieved for the query

```
retrieval systems AND multimedia AND images
```

retrieval systems

DOCID	POS
d1	10,25, 37
d2	23
d4	45,55
d7	3,56

multimedia

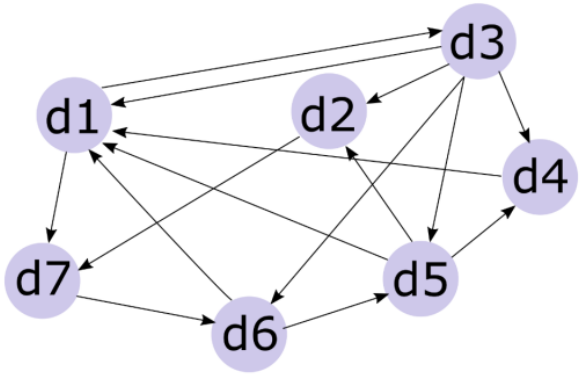
DOCID	POS
d4	18,20
d6	22,27
d7	5
d9	3,96, 127

images

DOCID	POS
d1	14,68
d3	29,73, 235
d4	11,145
d7	10

Ans : d4, d7

2. Below is the mini toy Web, which of the web pages would you expect to get the highest PageRank ?

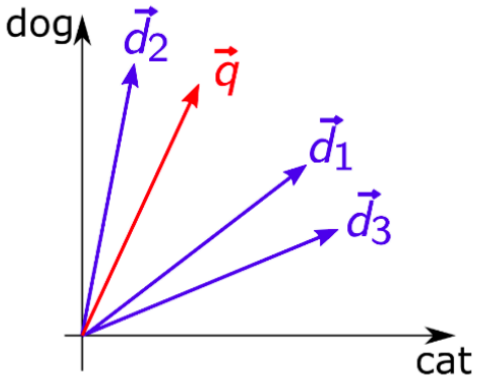


Ans : d1

3. What does PageRank calculate ?

Ans : The authority of a web page.

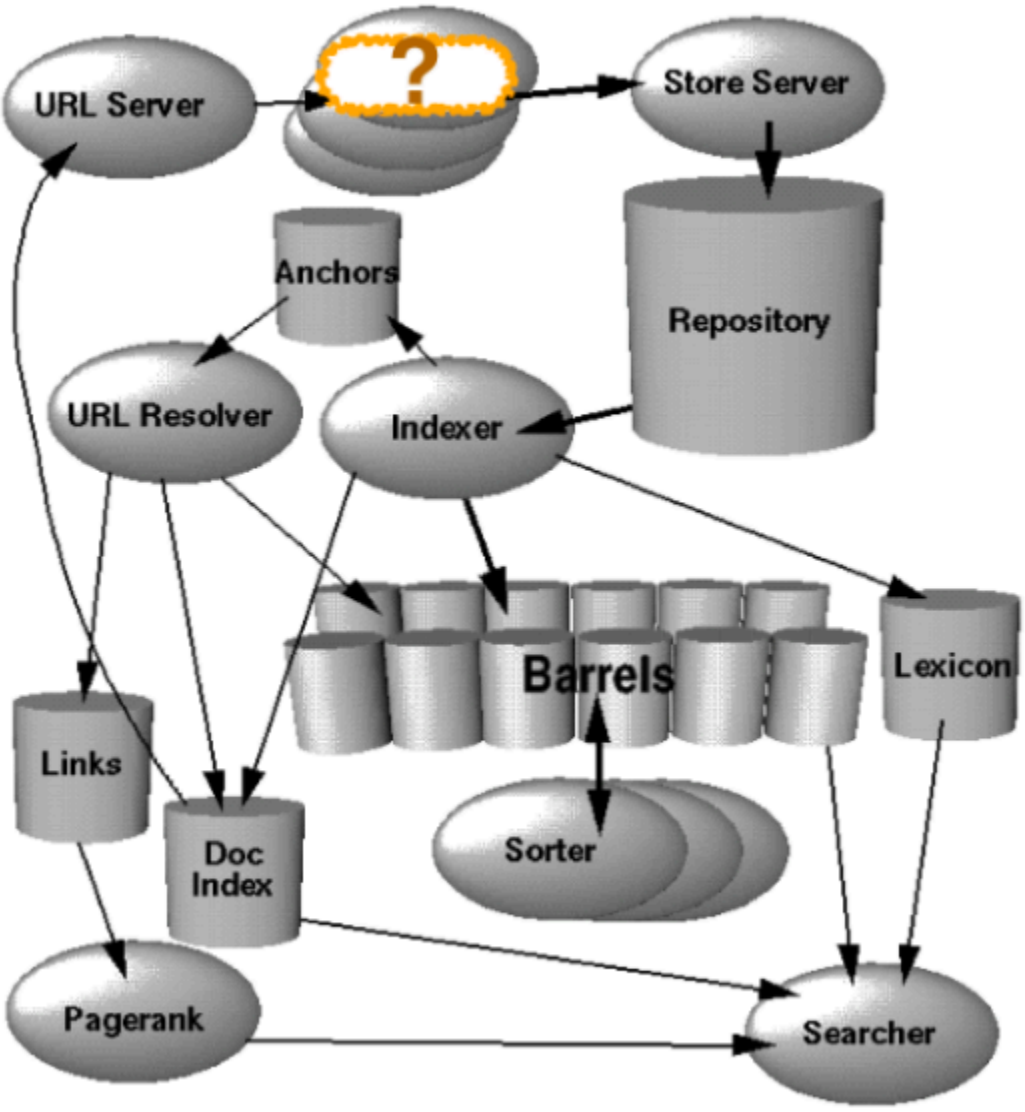
4. Consider the following situation in the image below, where we have a query vector q (in red) and 3 document vectors d1, d2, and d3 (in blue). Our toy index consists of the terms ‘cat’ and ‘dog’.
   How do you interpret this situation according to the vector space model ? Please choose one correct answer.



   Ans : d2 is mainly about dogs and only little about cats. d3 is mainly about cats but a bit about dogs, too. The user is interested in documents that are mainly about dogs but would be nice if cats are mentioned, too.

5. How does the HITS algorithm compute ?
   Ans : Hub and authority values for web pages.

6. Below is the google architecture as presented in the 1998 Brin and Page paper. What is the name of the missing component (orange question mark) ?



Ans : Crawler

Big Data Quiz

1. What is the purpose of lemmatization in NLP?

- ☐ Convert text into numerical format
- ☐ Reduce words to their root form
- ☐ Remove punctuation marks
- ☐ Identify part of speech tags

Ans : Reduce words to their root form

2. Which of the following is a common preprocessing step in NLP?
   - ☐ Compilation
   - ☐ Garbage Collection
   - ☐ Tokenization
   - ☐ Indexing

   Ans : Tokenization

3. Which of the following techniques is used to convert words into vectors?
   - ☐ Bag of Words
   - ☐ Word2Vec
   - ☐ TF-IDF
   - ☐ All of the above

   Ans : All of the above

4. Which scaling method is more fault-tolerant ?
   - ☐ Vertical Scaling
   - ☐ Horizontal Scaling
   - ☐ Both are equally fault-tolerant
   - ☐ Neither, as fault tolerance depends on backup strategies

   Ans : Horizontal scaling

5. What does a high positive correlation (close to +1) between two variables indicate?
   - ☐ One variable increases as the other decreases
   - ☐ One variable decreases as the other decreases
   - ☐ One variable increases as the other increases

- ☐ No relationship between the variables

Ans : One variable increases as the other increases

6. Which Python library is commonly used for EDA?
   - ☐ Scikit-learn
   - ☐ TensorFlow
   - ☐ Matplotlib
   - ☐ PyTorch

   Ans : Matplotlib

7. Which of the following is a categorical variable?
   - ☐ Age
   - ☐ Blood type
   - ☐ Height
   - ☐ Temperature

   Ans : Blood type

8. Which type of model is used for predicting a continuous numerical value?
   - ☐ Regression Model
   - ☐ Classification Model
   - ☐ Clustering Model
   - ☐ Reinforcement Model

   Ans : Regression Model