DS 5110: Introduction to Data Management and Processing (Fall 2024)

Sample Final Quiz

- You have 1.5 hours.
- The exam is closed book, closed notes.
- You are allowed to bring two letter-sized pages of notes (on both sides).
- Use of electronic devices is not allowed.
- Please read all instructions carefully.

Name: _____

NEU ID (optional): _____

1 Multiple-Choice Questions (50 points)

Circle **ALL the correct choices**: there may be more than one correct choice, but there is always at least one correct choice. **NO partial credit**: the set of all the correct answers must be checked. There are 10 multiple choice questions worth 5 points each.

1. When is a NoSQL database **not** a suitable replacement for a relational database?

   (a) When the data is unstructured and rapidly changing.

   (b) When the data cannot fit into the memory of a single machine.

   (c) When transactions are critically important for the application.

   (d) When dealing with big data applications that require high-speed data processing and analytics.

   (e) When strict schema enforcement is required for the application.

2. What are the advantages of using the BSON format over JSON?

   (a) BSON is optimized for data storage and retrieval.

   (b) BSON supports more data types than JSON.

   (c) BSON is more human readable than JSON.

   (d) BSON automatically compresses data to reduce storage space.

3. Suppose you have the following collection in MongoDB. The collection is named `contacts`.

```
{"_id" : 1, "region" : "NW1", "leads" : 1, "email" : "mlangley@co1.com"}
{"_id" : 2, "region" : "NW1", "leads" : 1, "email" : "jpicoult@co4.com"}
{"_id" : 3, "region" : "NW1", "leads" : 2, "email" : "zzz@company2.com"}
{"_id" : 4, "region" : "SE1", "leads" : 8, "email" : "mary@hssu.edu" }
{"_id" : 5, "region" : "SE2", "leads" : 4, "email" : "janet@col.edu"}
{"_id" : 6, "region" : "SE2", "leads" : 2, "email" : "bill@uni.edu"}
{"_id" : 7, "region" : "SE2", "leads" : 4, "email" : "iii@company1.com"}
{"_id" : 8, "region" : "SW1", "leads" : 1, "email" : "phil@co3.com"}
{"_id" : 9, "region" : "SW1", "leads" : 2, "email" : "thomas@company.com"}
{"_id" : 10, "region" : "SW2", "leads" : 2, "email" : "sjohnson@uchi.edu"}
{"_id" : 11, "region" : "SW2", "leads" : 5, "email" : "tsamuel@someco.com"}
```

   How many documents will be returned from the following query?

```
db.contacts.aggregate([
    {"$group": {"_id": "$region", "count": {$count: {}}}},
    {"$match": {"count": {"$gte": 3}}}
])
```

   (a) 0

   (b) 1

   (c) 2

   (d) 3

   (e) 4

4. Suppose you have the following collection called `orders` in MongoDB:

```
[
  { "_id": 1, "customer": "John Doe", "items": [{ "product": "Laptop", "price
    ": 1200 }, { "product": "Mouse", "price": 50 }] },
  { "_id": 2, "customer": "Jane Smith", "items": [{ "product": "Laptop", "
    price": 1200 }, { "product": "Keyboard", "price": 100 }] },
  { "_id": 3, "customer": "Mike Johnson", "items": [{ "product": "Smartphone",
     "price": 800 }] },
  { "_id": 4, "customer": "Sarah Brown", "items": [{ "product": "Smartphone",
    "price": 800 }, { "product": "Laptop", "price": 1200 }] }
]
```

Which MongoDB query finds all orders where the total price of items in the order is greater than $1000?

(a)
```
db.orders.find({ "items.price": { $sum: { $gt: 1000 }}})
```

(b)
```
db.orders.find({ "items": { $elemMatch: { "price": { $gt: 1000 }}}})
```

(c)
```
db.orders.aggregate([
        { $unwind: "$items" },
        { $group: { _id: "$_id", totalPrice: { $sum: "$items.price" }}},
        { $match: { totalPrice: { $gt: 1000 }}}
    ])
```

(d)
```
db.orders.aggregate([
        { $project: { totalPrice: { $sum: "$items.price" }}},
        { $match: { totalPrice: { $gt: 1000 }}}
    ])
```

5. In HDFS, what is the role of the NameNode?

   (a) Stores the Block IDs and locations of any given file in HDFS.
   (b) Executes file system namespace operations such as renaming files and directories.
   (c) Creates and deletes data blocks.
   (d) Chooses which DataNodes store the replicas of a given file.

6. Which of the following statements related to MapReduce are true?

   (a) Each mapper/reducer must generate the same number of output key/value pairs as it receives on the input.
   (b) The input to reducers is grouped by key.
   (c) The output type of keys generated by mappers must be of the same type as their input.
   (d) The output type of keys generated by mappers must be of the same input type of keys received by reducers.

7. In Spark, which of the following operations triggers the execution of RDD transformations?

   (a) `map()`
   (b) `reduce()`
   (c) `transform()`
   (d) `select()`
   (e) `collect()`

8. What does the generalization error of a machine learning model represent?

   (a) The error rate of the model on the test data.

   (b) The variance of the model's predictions on the test data.

   (c) The expected value of the error on new input.

   (d) The difference in error between the training and testing data.

9. You apply standard scaling to the training set using its mean and variance. Which statement about standardizing the test set is true?

   (a) The test set should be standardized using its own mean and variance.

   (b) The test set should be standardized using the training set's mean and variance.

   (c) The test set should be standardized using the overall mean and variance of both the training and test sets combined.

   (d) The test set should not be modified in any way.

10. In Scikit-Learn, which of the following statements is true about the `Pipeline` class?

    (a) It sequentially applies a series of data transformations followed by a final estimator.

    (b) The last estimator in the pipeline must have a `predict` method.

    (c) It automatically optimizes hyperparameters of each step in the pipeline.

    (d) It helps to prevent data leakage by ensuring that data transformations are only learned from the training set.

2 MongoDB (30 points)

You have a collection called `movies`. A sample document in the collection is shown below:

```
{'_id': ObjectId('573a1394f29313caabcdf639'),
 'title': 'Titanic',
 'released': datetime.datetime(1953, 7, 13, 0, 0),
 'plot': 'An unhappy married couple deal with their problems on board the ill-
    fated ship.',
 'genres': ['Drama', 'History', 'Romance'],
 'runtime': 98,
 'cast': ['Clifton Webb','Barbara Stanwyck','Robert Wagner','Audrey Dalton'],
 'directors': ['Jean Negulesco'],
 'writers': ['Charles Brackett', 'Walter Reisch', 'Richard L. Breen'],
 'imdb': {'rating': 7.3, 'votes': 4677},
 'reviews': [{'username': 'user1', 'comment': 'THIS MOVIE IS AMAZING!!! I love
    watching it!'},
    {'username': 'user2', 'comment': 'Knowing it is the 3rd highest grossing film
    of all time as of now, I can imagine why!'},
    {'username': 'user3', 'comment': 'This movie is a masterpiece, everything is
    perfectly and beautifully shot and well acted.'}]
}
```

Write the following queries in MongoDB (5 points for each query):

1. Find how many movies belong to the 'Romance' genre.

2. Find the titles and the ratings of the three movies with the highest IMDB rating.

3. Find the actor/actress who participated in the highest number of movies.

4. Find the average number of reviews written per movie. Note that some movies don't have any reviews.

5. Add an actor named "Samuel L. Jackson" to the movie "Pulp Fiction".

3 MapReduce (20 points)

For each of the following problems describe how would you solve it using MapReduce. Write the map and reduce tasks in pseudo-code.

1. The input is a list of housing data where each input record contains information about a single house: (address, city, state, zip, price). The output should be the average house price in each zip code.

2. The input contains two lists. The first list provides voter information for every registered voter: (voter-id, name, age, zip). The second list gives occupancy information: (zip, age, job). For each unique pair of zip and age values, the output should give a list of names and a list of jobs for people in that zip code with that age. If a particular zip/age pair appears in one input list but not the other, then that zip/age pair can appear in the output with an empty list of names or jobs, or you can omit it from the output entirely.