CS639.1 - Fall 2024

Instructor: Meenakshi Syamkumar

Final — 20%

(Last) Surname: _____ (First) Given name: _____

NetID (email): _____ @wisc.edu

Fill in these fields (left to right) on the scantron form (use #2 pencil):

1. LAST NAME (surname) and FIRST NAME (given name), fill in bubbles
2. IDENTIFICATION NUMBER is your Campus ID number, fill in bubbles
3. Under **F** of SPECIAL CODES, write **1** and fill in bubble **1**

If you miss step 3 above (or do it wrong), the system may not grade you against the correct answer key, and your grade will be no better than if you were to randomly guess on each question. So don't forget and double check it's correct!

You may only reference your note sheet. You may not use books, calculators, or other electronic devices during this exam. You may not sit near your friends or look at your neighbors during this exam. Please place your student ID face up on your desk. Turn off and put away portable electronics (including smart watches) now.

Use a #2 pencil to mark all answers. DO NOT USE PEN on the scantron.

When you're done, please hand in the exam and note sheet and your filled-in scantron form. The note sheet will not be returned.

(Blank Page)

Elasticsearch

Consider the below dataset containing information about 2982 books, indexed into elasticsearch for this section's questions.

```
[
  {
    "book_id": "B001",
    "title": "The Great Adventure",
    "author": "John Smith",
    "genre": "Fiction",
    "price": 12.99,
    "rating": 4.7,
    "reviews": [
      {
        "review_id": "R001",
        "reviewer": "Emily",
        "rating": 5,
        "comment": "Absolutely captivating story!"
      },
      {
        "review_id": "R002",
        "reviewer": "Michael",
        "rating": 4,
        "comment": "Engaging but a bit slow in the middle."
      }, ...
    ]
  },
  ... # info for the remaining 2981 books
]
```

1. How can we delete an existing index called "books"? You may assume that the variable `client` is appropriately initialized.

   A. `client.delete_index("books")`

   B. `client.indices.delete("books")`

   C. `client.delete_index(index="books")`

   D. `client.indices.remove("books")`

   E. `client.indices.delete(index="books")`

2. Which of the following queries will enable us to retrieve the titles of **all** the indexed documents?

A.
```
{
  "query": {
    "match": {
      "title": "all"
    }
  },
  "_source": ["title"],
  "size": 2982
}
```

B.
```
{
  "query": {
    "term": {
      "title": "all"
    }
  },
  "_source": ["title"],
  "size": 2982
}
```

C.
```
{
  "query": {
    "match_all": { }
  },
  "_source": ["title"]
}
```

D.
```
{
  "query": {
    "match_all": { }
    "_source": ["title"],
    "size": 2982
  },
}
```

E.
```
{
  "query": {
    "match_all": { }
  },
  "_source": ["title"],
  "size": 2982
}
```

3. What does the following query do?

```
{
    "query": {
        "match_all": {}
    },
    "size": 0,
    "aggs": {
        "books_aggregate": {
            "value_count": {
                "field": "price"
            }
        }
    }
}
```

A. counts the total number of documents that have value for the `price` field

B. calculates the sum of all values in the `price` field

C. calculates the average of all values in the `price` field

D. retrieves all documents in the index and includes the `price` field in the results

E. retrieves all documents where the `price` field is greater than zero

4. Which of the following aggregate query operators require the usage of `keyword` version of the search field?

A. `avg` B. `min` C. `max` D. `sum` **E.** `terms`

5. Which of the following queries would find all books in the title field that contain either "Science Fiction" or "Fantasy" with a higher score for "Science Fiction" (3.0)?

A.
```
{
  "query": {
    "simple_query_string": {
      "query": "‘Science Fiction’^3.0 ‘Fantasy’^3.0",
      "fields": ["title"],
      "default_operator": "or"
    }
  }
}
```

B.
```
{
  "query": {
    "simple_query_string": {
      "query": "‘Science Fiction’ ‘Fantasy’^3.0",
      "fields": ["title"],
      "default_operator": "or"
    }
  }
}
```

C.
```
{
  "query": {
    "simple_query_string": {
      "query": "‘Science Fiction’^3.0 ‘Fantasy’",
      "fields": ["title"],
      "default_operator": "and"
    }
  }
}
```

D.
```
{
  "query": {
    "simple_query_string": {
      "query": "‘Science Fiction’^3.0 ‘Fantasy’",
      "fields": ["title"],
      "default_operator": "or"
    }
  }
}
```

6. What does the following query do?

```
{
    "query": {
        "match": {
            "title": {
                "query": "Adventures in Wonderland",
                "fuzziness": "AUTO"
            }
        }
    }
}
```

A. searches for exact matches of the title "Adventures in Wonderland"

B. searches for titles that are similar to "Adventures in Wonderland" allowing for minor spelling variations

C. searches for title that contain the terms "Adventures", "in", "Wonderland" in any order

D. searches for reviews that exactly mention "Adventures in Wonderland"

E. searches for reviews that contain the terms "Adventures", "in", "Wonderland" in any order

7. What does the following query do?

```
"query": {
    "match": {
        "title": "Fantasy"
    }
},
"highlight": {
    "fields": {
        "title": {
            "pre_tags": ["<strong>"],
            "post_tags": ["</strong>"]
        }
    }
}
```

A. returns all books whose `title` contains "fantasy", highlighting the `title` in bold

B. returns all books, highlighting only the results whose `title` contain "fantasy" in bold

C. returns all books whose `title` contains "fantasy", highlighting the entire search result in bold

D. returns all books, highlighting any occurrence of the term "fantasy" in bold

MongoDB

Consider the below `books` collection for this section's questions. You may assume that the variable `db` is appropriately initialized.

```
[
  {
    "book_id": "B001",
    "title": "The Great Adventure",
    "author": "John Smith",
    "genre": "Fiction",
    "price": 12.99,
    "rating": 4.7,
    "reviews": [
      {
        "review_id": "R001",
        "reviewer": "Emily",
        "rating": 5,
        "comment": "Absolutely captivating story!",
        "date": "November 1st 2020"
      },
      {
        "review_id": "R002",
        "reviewer": "Michael",
        "rating": 4,
        "comment": "Engaging but a bit slow in the middle.",
        "date": "April 14th 2023"
      }, ...
    ]
  },
  ...
]
```

8. Which of the following queries would return all books that have at least 2 reviews?

   A. `db.books.find({ "$expr": { "$lt": [{ "$size": "$reviews" }, 2] } })`

   B. `db.books.find({ "reviews":{ "$size": 2 } })`

   C. `db.books.find({ "$expr": { "$gte": [{ "$size": "$reviews" }, 2] } })`
   .

   D. `db.books.find({ "reviews": { "$exists": true } })`

9. Which of the following queries would return the latest five reviews for each book. Assume that the `reviews` are ordered based on ascending order of `date`.

   A. `db.books.find({}, { "reviews": { "$slice": 5 } })`

   B. `db.books.find({}, { "reviews": { "$slice": -5 } })`

   C. `db.books.find({}, { "reviews": { "$slice": [-5, 5] } })`

   D. `db.books.find({ "reviews": { "$slice": 5 } })`

   E. `db.books.find({ "reviews": { "$slice": -5 } })`

10. What does the following code do?

```
pipeline = [
    {
        "$group": {
            "_id": "$genre",
            "average_price": {"$avg": "$price"}
        }
    },
    {
        "$project": {
            "_id": 1,
            "average_price": {"$round": ["$average_price", 2]}
        }
    }
]
avg_price = list(db.books.aggregate(pipeline))
avg_price
```

   A. calculates the total price of books for each genre, rounded to 2 decimal places

   B. calculates the total price of books for each genre

   C. calculates the total price of books, rounded to 2 decimal places

   D. calculates the average price of books, rounded to 2 decimal places

   E. calculates the average price of books for each genre, rounded to 2 decimal places

11. Which of the following pipeline operators performs join operation?

   A. `$group` B. `$project` C. `$match` **D.** `$lookup` E. `$unwind`

12. Which of the following pipeline stages are **not** required to answer this question: Find the top 3 authors who have written the most books.

   A. `$match` B. `$group` C. `$sort` D. `$limit`

Other topics

13. Which of the following commands will take us to the home directory in Linux?

    A. `cd .` B. `cd HOME` C. `cd ..` D. `cd +` **E.** `cd ~`

14. Given below are the schemas of SQL tables:

    ```
    Books:
    | book_id | title | author | genre | price | rating |

    Reviews:
    |review_id | book_id | reviewer | rating | comment | date |
    ```

    What does the below SQL query compute?

    ```
    SELECT
        b.book_id, b.title,
        COUNT(r.review_id) AS totalReviews,
        RANK() OVER (ORDER BY COUNT(r.review_id) DESC, b.title ASC) AS review_rank
    FROM books b
    JOIN reviews r ON b.book_id = r.book_id
    GROUP BY b.book_id, b.title
    ```

    A. computes the rank of each book based on the number of reviews, with ties broken by the title in ascending order

    B. computes the rank of each book based on the average rating of its reviews, with ties broken by the title in descending order

    C. computes the total number of reviews for each book, without ranking them

    D. computes the total number of reviews for each book, with rank based on the number of reviews, but without considering the title for tie-breaking

15. What does the differencing technique enable us to do in time-series analysis?

    A. determine `p` value for ARIMA

    B. determine `q` value for ARIMA

    C. transform a non-stationary series into a stationary series

    D. compute moving average of the data points

16. While fine-tuning an LLM model, training and validation loss between epochs should _____.

    A. increase **B. decrease**

17. Which normalization type requires partial dependencies to be eliminated?

A. Denormalized B. 1NF **C. 2NF** D. 3NF

18. In LLM models, which process enables us to transform a sequence of numerical tokens into human-readable text?

A. quantization B. tokenization C. encoding **D. decoding**

19. How can we extract the date from a `pandas DataFrame` column `events["event_type"]` that has datetime format?

A. `events["event_type"].date`

B. `events["event_type"].dt.date`

C. `events["event_type"].to_date()`

D. `events["event_type"].extract_date()`

20. Which of the following forecasting algorithms performs leaf-wise growth, where the tree expands its most significant leaf nodes first?

A. EMA B. Linear regression C. ARIMA D. XGBoost **E. LightGBM**

21. Which message delivery semantics can incur data loss?

A. exactly-once semantics

B. at-least-once semantics

C. at-most-once semantics

D. no delivery semantics

22. Which of the following data models expects data to be fully normalized?

A. Inmon data model B. Kimball data model C. Data vault model

23. What is transitive dependency?

A. non-key column depends on another non-key column

B. non-key column depends on part of the primary key column

C. primary key column depends on non-key column

D. primary key column depends on another primary key column

24. Which of the following Streamlit functions enables us to configure a drop-down box for the user to provide input from a set of values?

A. `st.selectbox` B. `st.text_input` C. `st.slider` D. `st.columns`

25. Which of the following `pandas` methods enables us to determine the count or frequency corresponding to each unique value within a `series`?

A. `is_unique` B. `nunique` C. `unique` **D.** `value_counts`

26. What is the purpose of data interpolation?

A. to remove outliers from the data

B. to remove rows containing missing values from the data

C. to fill missing values with estimated values

D. to sort data in ascending order

27. Which of the following correlation values means a perfect correlation between two variables?

A. -4 B. -3 C. -2 **D. -1** E. 0

28. Why should we ignore the first data point in ACF and PACF plots?

A. it is often noisy and should be eliminated

B. it is used as a benchmark for evaluating other lags

C. it always has a value of 0, making it irrelevant

D. it always has a value of 1, as it represents the autocorrelation of the series with itself

29. Which of the following will enable us to redirect just `stdout` to a file named `out.txt`?

A. `> out.txt` B. `2> out.txt` C. `&> out.txt` D. `| out.txt`

30. Which SQL clause enables us to perform projection?

A. `SELECT` B. `FROM` C. `WHERE` D. `GROUP BY` E. `HAVING`

31. In the context of data ingestion, what does pull type of ingestion refer to?

A. source system sending data to a target

B. target reading data from a source

C. target periodically checking source for any changes and then performing pull accordingly

32. Which of the following SQL window functions enables us to compute ranks with ties? **QUESTION DROPPED DUE TO TWO CORRECT ANSWERS**.

A. `RANK` B. `DENSE_RANK` C. `ROW_NUMBER`

33. What does decreasing `max_new_tokens` do to a pre-trained LLM model's output?

A. helps overcome hallucination effect

B. increases the length of the generated text

C. decreases the length of the generated text

D. helps improve accuracy of the generated text

34. Which of the following `pandas` functions enables us to join data from two `DataFrames`?

A. `concat` B. `join` C. `combine` D. `append` **E.** `merge`

35. Which of the following arguments for the `resample` method will enable us to compute monthly statistics?

A. `'D'` **B.** `'ME'` C. `'Q'` D. `'W'` E. `'Y'`

(Blank Page)