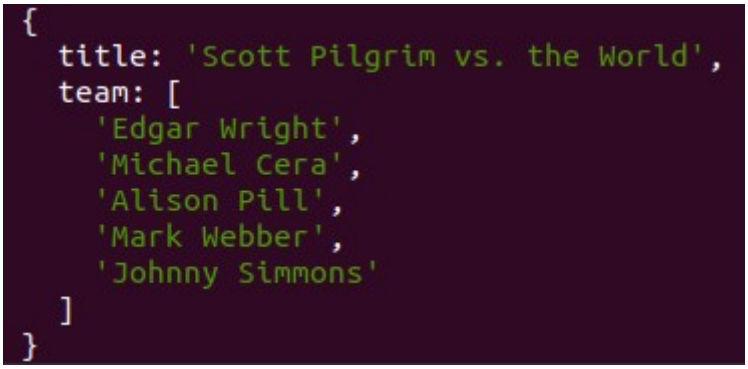
Data Engineering Lab 2021 (CS5103) Final Exam

I.

Use the provided mongoDB collections ‘movies’ and ‘comments’. Submit screenshots of queries.

1. Display titles of movies for which the user with email id 'megan_richards@fakegmail.com' wrote comments for in the years 2007 and 2014.
2. Display the title and date of release of movies whose titles start with 'aa' and end with 'm' (case-insensitive).
3. Combine the ‘directors’ and ‘cast’ fields of the movies released in 2010 into a new field 'team' and display them along with the title of the movies.

In the screenshot below showing the expected output, ‘Edgar Wright’ is the director of the movie and the rest form the cast:



4. Display the titles and plots of all 'Documentary' films with India in their 'countries' field.

II.

1. Read 'shakespeare-hamlet' from the Gutenberg corpus in raw form (as a single string). Find all words that end in 'ed' and display:
    1. The complete word.
    2. The same word with its ending 'ed' stripped out.
    3. The [start, end) indices of the match in the input string.

    Sample output:

    reminded
    remind
    (4, 12)

2. Find and display the stopwords that are present in 'shakespeare-macbeth' text in the Gutenberg corpus.
3. Read text 'chesterton-ball' from the Gutenberg corpus and
    1. Find and display the the most frequent, second most frequent, least frequent and second least frequent words.

2. Replace the most frequent word with the least frequent one and the second most frequent word with the second least frequent one and display the five most common words in the new frequency distribution.

III.

Load housing price dataset using pandas Dataframe. Consider the last column as target feature and remaining columns as input features.

1 Do following data cleaning and preprocessing steps:
- 1.a Handling missing values :
  - a.i Find out the columns containing missing values.
  - a.ii Remove all columns which contain more than 25% missing values. Display the eliminated column names.
  - a.iii Impute the missing values with either previous value , next value or mean value.
- 1.b Anomaly detection and removal :
  - b.i Visualise each column with a box plot or histogram . And observe which columns have outliers.
  - b.ii Removing outliers of observed columns using InterQuartile range.
- 1.c Feature Scaling :
  - c.i Implement feature scaling using either Normalisation or Standardization.

2 Modeling :
- 2.a Divide the dataset into training and testing dataset.
- 2.b Implement Linear regression on training data.

3 Evaluation :
- 3.a Evaluate the model using regression evaluation technique RMSE on test data.

4 From the above trained model print the top 3 important features.

IV.

Create a database table using the below schema. And Insert all records that are given in the csv file. Implement below queries and for each query upload a screenshot that contains the query as well as query output.

Schema for Student table :

{ *StudentID varchar primary key,*
StudentName varchar,
StudentAge int,
StudentDept varchar,
StudentCgpa float }

1 Write query for retrieving toppers of each department in descending order and student cgpa should be more than 5.0.

2 Write a trigger that shouldn't allow inserting students data having *cgpa* greater than 10.0 and less than 3.0.

3 Write a procedure that can delete records of students whose *cgpa* is less than average of department *cgpa* .