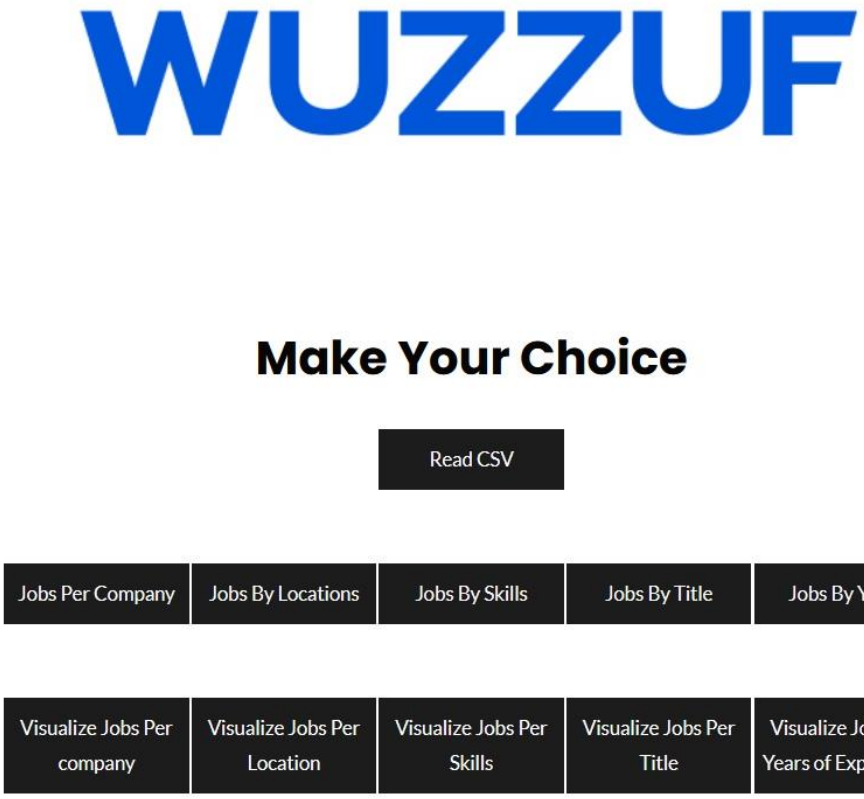

Wuzzuf Spark Spring Boot

(Team 3)

Project Link :-

- https://github.com/FadyNasser/Wuzzuf_Spark_Spring_Boot

Team Members :-

- Fady Nasser Fawzy
- Saieed Osama Saieed
- Omar Ayman Abdelsamad

Overview

- A Web Application of Apache Spark using Spring Boot MVC.
- Link :- http://localhost:8080/Wuzzuf

Goals

1. Load Wuzzuf Dataset to perform Data science operations on it.
2. Build an Apache Spark Application to deal with this dataset.
3. Integrate The Apache Spark with Spring Boot to build a web application that handles this dataset.
4. Preview some insights obtained from the dataset.

Specifications

- Spring is a very popular Java-based framework for building web and enterprise applications.
- Spring framework provides flexibility to configure beans/objects in multiple ways such as XML, Annotations, and JavaConfig.
- Spring's dependency injection approach encourages writing testable code
- Spring Boot is basically an extension of the Spring framework which eliminated the boilerplate configurations required for setting up a Spring application.
- Apache Spark is an open-source distributed general-purpose cluster-computing framework.
- Java JDK Version 1.8.
- DataSet Link :- https://www.kaggle.com/omarhanyy/wuzzuf-jobs

Milestones

I. Wuzzuf Spark Only

- A Spark Application used to read wuzzuf dataset and then Convert it to data frame and display some of them.
- Display structure and summary of the data.
- Clean the data (null, duplications).
- Count the jobs for each company and display that in order and Show them in a pie chart.
- Get the most popular job titles and Show them in a bar chart.
- Get the most popular areas and Show them in a bar chart).
- Print skills one by one and how many each repeated and order the output to find out the most important skills required.
- Factorize the YearsExp feature and convert it to numbers in a new column.
- Apply Linear Regression to job titles and companies with respect to the factorized YearExp.
- Apply K-means for job titles and companies.

II. Wuzzuf Spark Spring Boot

- A Web Application of Apache Spark using Spring Boot MVC.
- Integrated the Wuzzuf spark application with spring boot to have a web application that perform the tasks separately and show them one then the other.
- Interface made using HTML and CSS which is responsible for sending and getting requests from the web application.

Project Modules and Classes

I. Main Class

- A Class responsible for starting the spring application and directing the app to the index.html home page found in “src/resources/templates/index.html”.

II. Application Config Class

- Contains WebMvcConfig Class which is responsible for handling images resources.
- Contains ApplicationConfig Class which is responsible for initializing the web app.

III. Spark Controller Class

- Which is the main class of our project that starts spark session then handles every tasks according to the request it gets, which contains the following :-
    1. Wuzzuf_ReadCSV :- Responsible for reading the dataset, removing nulls values and duplicates then print some of it with the dataset schema and summary.
    2. Wuzzuf_JobsByCompany :- Responsible for printing each job with its count and generate a pie and bar chart for them.
    3. JobsPerCompanyImg :- Responsible for printing the Pie chart of jobs.
    4. Wuzzuf_JobsByCompany :- Responsible for printing each job with its count and generate a pie and bar chart for them.
    5. JobsPerCompanyImg :- Responsible for printing the Pie chart of jobs.

6. JobsPerCompanyWithoutConfImg :- Responsible for printing the Pie chart of jobs without “Confidential” Record for more efficient data representation.
7. Wuzzuf_JobsByTitle :- Responsible for printing each Title with its count and generate a bar chart for them.
8. JobsPerTitleImg :- Responsible for printing the bar chart of titles.
9. Wuzzuf_JobsByLocations :- Responsible for printing each Location with its count and generate a bar chart for them.
10. JobsPerLocationImg :- Responsible for printing the bar chart of Locations.
11. Wuzzuf_JobsBySkillss :- Responsible for printing each Skill with its count and generate a bar chart for them.
12. JobsPerSkillsImg :- Responsible for printing the bar chart of Skills.
13. Wuzzuf_JobsByYears :- Responsible for transforming “YearExp” Column to integer by splitting the record or “YearExp” and get first digit then group by them to print each Year with its count and generate a bar chart for them. After that, it split the dataset to lists of dataset rows according to the year or experiences and print each group of years separately.
14. JobsPerYearsofExperiences :- Responsible for printing the bar chart of Years of experiences.
15. Wuzzuf_MachineLearning :-
    - Works only on the Wuzzuf-Spark-only Project as vector assembler conflicts with Spring Boot
    - It is responsible for Using Linear Regression to predict Titles and Company with respect to Years of Experiences and also Using K-Means to Predict Titles and Companies.
    - Accuracy isn’t working fine since we used to transform titles and companies to integer vectors by adding an ID column to classify according to it.

16. BarChart_Data Function :- Used to Plot Bar charts for given lists and save them as jpg to be loaded by the web app.
17. PieChart_Data Function :- Used to Plot Pie charts for given lists and save them as jpg to be loaded by the web app.
18. ML_Prediction_Title :- Linear Regression model for predicting Title with respect to Years of Experiences.
19. ML_Prediction_Comapany :- Linear Regression model for predicting Company with respect to Years of Experiences.
20. KMeans_Prediction_Title :- KMeans Model used to predict Title by making clusters with same size of data.
21. KMeans_Prediction_Company :- KMeans Model used to predict Company by making clusters with same size of data.

IV. Webapp Directory

- Contains Images that are shown by the Sparkcontroller requests.

V. HTML Page

- Found in “src/resource/templates/index.html” which contains the code of the interface of the web app.

Steps To Run

I. Wuzzuf Spark Spring Boot

1. Download the project.
2. Run the application.
3. Browser Url :- http://localhost:8080/Wuzzuf
4. Click on the desired button to print the data.
5. Press the browser back to navigate to another button.

II. Wuzzuf Spark Only

1. Download the project.
2. Run the application.
3. Wait for the output to come one after the other till the program terminates.