3803ICT

Big Data Analysis

Assignment Specifications

Trimester 1 - 2020

Instructions

- **Due**: Friday, 22 May 2020 at 11:59pm
- **Marks**: 30% of your overall grade
- **Data**: https://drive.google.com/file/d/1DQPRoxIPdOyq_9Fg95hyj4taK2r0lbrd/view
- **Group Work**: You must complete this assignment in a group of maximum 2 students.

Overview

In this assignment, you will need to apply data analytics, using the tools introduced during the labs. You are required to study the SEEK job market data and analyze. The assignment consists of 3 parts. In the first part, you will need to understand data characteristics using data preparation and preprocessing techniques. In the second part, you will perform various data analysis techniques, including exploratory, statistical, and predictive ones. In the third part, you will need to evaluate your findings and determine appropriate future actions.

Part 1 –Data Preparation and Preprocessing [8 points]

- The primary dataset that we would like to use is the job market dataset which is provided in CSV format (data.csv).
- Perform data preparation and preprocessing for your analysis
- Submit your Jupyter notebook in your Github repository

1) Describe the dataset.
   For example:
   - What are the categories/domains of the dataset?
   - What is the dataset size of each variation?
   - What is dataset structure/format?
   - What are attributes/features of review data you are going to use?
   - What are attributes/features of product data you are going to use?
   - Which parts of the dataset will you use or all of them?

   [1-2 paragraphs, 3 points]
2) Describe the steps you used for data preparation and preprocessing.
   For example:
   - How do you load the data using Pandas?
   - How do you normalize the data?
   - How do you clean the data?

   [2-3 paragraphs, 4 points]
3) What is your hypothesis (expectation) about the analysis outcome?
   [1-2 paragraphs, 1 point]

Part 2 – Data Analysis and Interpretation [17 points]

- Perform exploratory data analysis
- Perform statistical data analysis
- Perform predictive data analysis
- Submit your Jupyter notebook in your Github repository

1) Study the job metadata. Extract the relevant information to describe the job's attributes.
   For example:
   - What is the sector, sub-sector of each job?
   - Where is the location of the job?
   - Which is the range of salaries for each job?

   [1-2 paragraphs, 5 points]
2) Study the market by locations.
   For example:
   - What is the market size in each city? Which are the hottest job sectors in each city?
   - Which range of salary is common in each city? Where are the employees more well-paid?
   - Can you detect the pattern of posting: e.g. are more jobs posted at the beginning of month?

   [1-2 paragraphs, 4 points]
3) Study the market by sectors.
   For example:
   - Which sectors keep the highest market share?
   - In each sector, which sub-sectors are the main spotlights?
   - What is the salary range for each sector/sub-sector? Can you compare salary range between sectors/subsectors?
   - What is the trending of market i.e. if a high school student ask you which subject should he/she learn in the university (to guarantee a job in a future), what is your advice?
   - Can you detect which skills are required in each sector?

   [1-2 paragraphs, 4 points]
4) Visualize the results on an interactive web page
   For example:

- Trend analysis: visualize number of jobs by locations, by sectors, etc.
- Compare between locations or sectors about the number of jobs, the salary, etc.
- Present the necessary skills by sectors, by subsectors.

[1-2 paragraphs, 3 points]

5) Using PySpark for your data analysis [1 point]

Part 3 - Evaluation [5 points]

1) What are the findings of your data analytics?
   [2-3 paragraphs, 2 points]
2) What actions for balancing the markets do you suggest based on your findings? [1-2 paragraphs, 1 point]
3) How could you refine your data analytics?
   For example:
   - Could you use different data sources?
   - Could you choose different parameters?
   - Could you choose other techniques?
   - Can you think of ways to obtain more relevant data?

   [1-2 paragraphs, 1 point]
4) Are there any implications for employers and employees based on the findings you obtained? Justify your answer.
   [1-2 paragraphs, 1 point]
5) Present and visualize your data story on an online Web page [OPTIONAL - up to 5 bonus points]
   - Publish your website on Google Cloud (or github.io).