Students with difficulty in meeting the deadline because of illness, etc. must apply for an **assignment extension** in CloudDeakin no later than 12:00pm on 21/05/2021 (Friday).

- This is a group work for group with up to 3 members. If you choose to work on it individually, please seek approval from unit chair via email.
- There are folders for this task on CloudDeakin, please enrol into the group (2 or 3 members) before 15/05/2021 (12:00am):

  2021 Assessment Task 2 (1-member Group) for students with approval of working alone; Approval required;

  2021 Assessment Task 2 (2-member Group) for groups of 2 members; Self-enrollment required;

  2021 Assessment Task 2 (3-member Group) for groups of 3 members; Self-enrollment required.

- Please form the group first, and then self-enrol into the appropriate group before 15/05/2021 (12:00am).

Instructions

Six files are provided for this assessment task:

`HTWebLog_p1.zip` The compressed `zip` file is for Part I of this assessment task, and it is a sample of *Hotel TULIP* Web log dataset, which contains the web access log information from 11/2006 to 02/2007. [1].

`Citation2003-2021.csv` This `CSV` file is for Part II of this assessment task, and the file structure is provided.

`Search-results.csv` This `CSV` file is for Part II of this assessment task, and the file structure is provided.

`SIT742Task2.ipynb` This is the notebook file for the `Python` code in `ipynb`, and the latest notebook is also released in `SIT742Task2.ipynb`.

Web log This code snippet contains all the coding requirements and also hints for Part I of this assessment task.

Predictive Aanalysis This code snippet contains all the coding requirements and also hints is for Part II of this assessment task.

You will need to complete the code in the notebook and make it run-able. The results of running the notebook will help you to develop your report, as well as generate the required files: `Citation2003-2021.csv` and `Search-results.csv`.

`SIT742Task2-Report-Template.docx` This is the `Word` template for your report `SIT742Task2-Report.pdf`.

What to Submit?

You are required to submit the following completed files to the corresponding *Assignment* (Dropbox) in CloudDeakin:

`SIT742Task2.ipynb` The completed notebook with all the run-able code on all requirements.

`SIT742Task2-Report.pdf` Your report for the both Part I and Part II of this assessment task.

[1] This file is exclusively for `SIT742` educational purpose only. You are not allowed to further distribute it.

`Citation2003-2021.csv` The completed citation information as `CSV` file, sorted by year.

`Search-results.csv` The completed parameter grid search result as `CSV` file.

Part I
Data Analytic — *Web Log Data* (20 marks)

Here is the hypothetical background:

> *Hotel TULIP* (a hypothetical organisation) is a five star hotel that locates in Australia. It is a very special hotel with an equally special purpose: Not only does it embody all the creative energy and spirit of TULIP-Lab, it's a "learning environment" on which the tourism and hospitality students are trained for future hoteliers.
>
> In the past two decades, the Web server of *Hotel TULIP* has logged all the web traffic to the hotel website, and stored large amount of data related to the use of various web pages. The hotel's CIO, Dr *Bear Guts* (not *Bill Gates*!), believes that those log files are great resources to help their *Information Technology Division* improve their potential customers' online experience, and help their *Market Promotion Division* to identify potential customers and their behaviour patterns. Hence, *Hotel TULIP* would like you *Group-SIT742* (a hypothetical data analytics group with up to 3 data analysers) to analyse web log files and discover user accessing patterns of different web pages.
>
> The Web server is using *Microsoft Internet Information Service* (IIS), and the Web log format can be found at: `https://msdn.microsoft.com/en-us/library/ms525807(v=vs.90).aspx`

Task Description

This task requires you to develop a data analysis report for the provided *Hotel TULIP* Web logs.

Without exploration or further analysis, 'raw' Web log data hardly reveals any insightful information. In this part, you are required to complete the `Python` code snippets to generate suitable numeric and visual description in the *Hotel TULIP* Web log dataset based on the detailed requirements in `SIT742Task2.ipynb`, and develop the report `SIT742Task2-Report.pdf` to summarise the data analytic results. The detailed requirements can also be found in the notebook `SIT742Task2.ipynb`, here we summarise them as follows:

1 Data ETL (4 marks)

1.1 Load Data (2 marks)

Load data from files. In order to reduce the processing time, we will remove missing values, and select 30% of total data for the following tasks.

Code
- Remove missing values. For the columns, if the column is with 15% NAs, you need to remove that column. Then, for the rows, if there are any NAs in that row, you need to remove that row (requests)
- Randomly select 30% of the total data into a new `dataframe weblog_df`.

Report
- Please show the number of requests in `weblog_df`.

1.2 Feature Selection (2 marks)

Code Select '`cs_method`', '`cs_ip`', '`cs_uri_stem`', '`cs(Referer)`' as features and '`sc_status`' as the class label into a new `dataframe ml_df` for following Machine Learning Tasks.

Report
- Data Description of `ml_df`.
- Show the top 5 rows of `ml_df`.

2 Unsupervised learning (4 marks)

You are required to complete this part using `sklearn`.

Code
- Perform unsupervised learning on `ml_df` with *K Means*.

Report
- Visualization of '`KMeans`' performance using the elbow plot , with a varying `K` from 2 to 10.
- What is the best `K` for this dataset?

3 Supervised learning (8 marks)

You are required to complete this part using `PySpark` packages.

3.1 Data Preparation

Prepare the data for supervised learning by completing the code.

3.2 Logistic Regression (4 marks)

Code
- Perform supervised learning on `ml_df` with *Logistic Regression*.
- Evaluate the classification performance using confusion matrix including TP, TN, FP, FN;
- Evaluate the classification performance using *Precision*, *Recall* and *F1* score.

Report
- Show the classification result using confusion matrix.
- Evaluate the classification performance using confusion matrix including TP, TN, FP, FN,
- Evaluate the classification performance using *Precision*, *Recall* and *F1* score.

3.3 K-fold Cross-Validation (4 marks)

You are required to use *K-fold cross validation* to find the best hyper-parameter set where `K = 2`.

Code
- Implement *K-fold cross validation* for three (any three) classification models.

Report
- Your code design and running results.
- Your findings on the classification model or its hyper-parameters based on cross-validation results (Best results).

4 Association Rule Mining (4 marks)

You are required to complete this part using suitable package for association rule mining.

Code
- Analyze the dataset using association rule mining;
- Choose suitable threshold for *confidence*, *support* and/or other parameters.

Report
- Your code design and running results.
- Your findings on the association rule mining results.

Part II
Data Analytic — Prediction (8 marks)

Google Scholar is a web service that indexes the metadata of research articles on many scientists. Majority of computer scientists use *Google scholar* to track their publications and research development. Therefore, the web crawling on *Google Scholar* can provide the citation information on all professors with a public *Google Scholar* profile. After the crawling, the prediction could be conducted to predict the future citation numbers such as *citation all*.

Task Description

In 2021, to better introduce and understand the research works on the professors, the university wants to perform the citation prediction for individual professors. You are required to implement a web crawler to crawl the *citation* information for A/Professor Gang Li from 2003 to 2021 (inclusive), and also conduct several prediction as required. You will need to make sure that the web crawling code and prediction code meets the requirements. You are free to use any `Python` package for Web crawling and prediction by finishing the following tasks.

1. Crawl the *citation* information for A/Professor Gang Li from 2003 to 2021.
2. Train Arima on *citation* information from 2003 to 2017, and predict the 2018, 2019 and 2020 *citation* information. You need to draw the line plot [2] to show the predicted citation for comparison (more details in below sections).
3. Conduct the grid search on Arima parameters (p, d and q) to select the best parameter values and then use them to predict the *citation* information from 2021 to 2022. You also need to draw the prediction for comparison (more details in below sections).

5 A/Professor Gang Li *citation* Information Extraction

You will need to import the suitable (or your chosen) web crawling library and use the corresponding library to crawl the year 2003 to year 2021 *citation* information (19 years) for A/Professor Gang Li's google scholar profile page: `https://scholar.google.com/citations?user=dqwjm-0AAAAJ`. Eg: *citation* on year 2020 is 839 and *citation* on year 2021 is 228 [3].

5.1 Crawl and Generate the citation `dataframe` (1 mark)

The code must contain the necessary web crawling steps and necessary data saving steps. The results of the code running will generate the `citation2003-2021.csv`. The `citation2003-2021.csv` will contain the year column and citations column. Data extraction without web crawling steps in the code will incur 0 mark.

6 Train Arima to predict the 2018 to 2020 citation

In this part, you need to train the Arima, perform the prediction and also evaluation.

6.1 Train Arima Model (1 mark)

You will need to use the crawled `citation2003-2021.csv` and then perform the Arima training with parameter of $p = 1$, $q = 1$ and $d = 1$ on data from 2003 to 2017 (15 years).

[2] for some coding example, please check `https://stackabuse.com/matplotlib-line-plot-tutorial-and-examples/`

[3] Hint: In the right corner of Google profile page, there is a hyperlink `VIEW ALL`. By clicking the hyperlink, you could see all the citations from 2003 to 2019

6.2 Predicting the citation and Calculate the RMSE (1 mark)

Then you will need to use the trained ARIMA model to predict the citation on year 2018, 2019 and 2020. You will need to perform the evaluation by comparing the predicted citation from 2018 to 2020 with the true citation from 2018 to 2020 and calculate the *root mean square error* (RMSE).

6.3 Visualization for comparison (1 mark)

You will also need to use *matplotlib* to draw the line plot with training data from 2013 to 2017, the testing true value, the prediction value and also the confidence interval.

Note

You will need to complete the notebook code, and insert the related self-written code and required results into the corresponding place of the report `SIT742Task2-Report.pdf`.

7 Parameter selection and 2021-2022 Prediction

In this part, you will need to conduct the grid search with ARIMA and select the best parameter values to predict the citations on 2021 and 2022.

7.1 Grid Search (2 mark)

You will need to run the grid search for parameters from the range $p = [1, 2]$, $q = [1, 2]$, $d = [1, 2]$ with training data (year 2003 to 2017) and testing data (year 2018 to 2020). The result of the search on each parameter combination (eg: p=1, q=1, d=1) will need to be stored in the `search-results.csv`. The `search-results.csv` will have the column of "RMSE" and column "parameter-set".

7.2 Select the best parameter values and Predict for 2021 and 2022 (2 marks)

You will need to perform the training with ARIMA on data from 2003 to 2020 with best parameter values you have found above, and then conduct the prediction for year 2021 and 2022. You will also need to use `matplotlib` to draw the line plot with training data from 2013 to 2020, the predictions 2021 to 2022 together with their confidence interval.

Note

You will need to complete the notebook and insert the related self-written code and required results into the corresponding place of the report `SIT742Task2-Report.pdf`.

Part III
Self Reflection - Essay (12 marks)

8 Self Reflection Essay

Based on your experience with the assessment tasks, you are required to write an essay with 1200-2000 words to reflect your understanding and thoughts on the Big data, which should include the following information:

1. What are the Python packages that you find useful in manipulating and analyzing Big data? You can briefly analyze their advantages and disadvantages;

2. What are the Big data platforms that can help storing, retrieving and analyzing the big data? What are their advantages and disadvantages?
3. Compare and contract the Python data analytical packages and their Spark packages.
4. What are your opinions on the privacy issues in the Big data era? Any example to further illustrate the risks?
5. What are the methods you think could help to solve the privacy issues on big data? Please list any successful implemented method.
6. Any other thoughts about data science, or suggestions to future students (or teaching team) about this unit.

Referencing should be in Harvard style, and more information about essay writing can be found at:

- `https://www.deakin.edu.au/students/studying/study-support/academic-skills/essay-writing`, and
- `https://www.deakin.edu.au/students/studying/study-support/academic-skills/reflective-writing`.
- `https://www.deakin.edu.au/students/studying/study-support/referencing/referencing-explained/introduction`