## **Declaration on Plagiarism**

*This form must be filled in and completed by the student(s) submitting an assignment*

|  |  |
| --- | --- |
| **Name:** | Prateek Sakaray - GC (19210825)  Omkar Kulkarni  (19210641)  Vinit Bhosale (19210115)  Shivam Moray (19211045)  Anuja Kaulgikar (19211052)  Chethana Prasad HG (19210707) |
| **Programme Code:** | MSc in Computing |
| **Module Code:** | CA675 |
| **Title:** | Airport and aircraft evaluation system |
| **Group Number:** | 8 |
| **Submission Date:** | 14/12/2019 |
| **Professor:** | Dr Long Cheng |

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at <http://www.dcu.ie/info/regulations/plagiarism.shtml>, <https://www4.dcu.ie/students/az/plagiarism> and/or recommended in the assignment guidelines

Name: Prateek Sakaray (GC) Date: 14/12/2019

**Assignment 2: Airport and Aircraft Evaluation System**

**Introduction and Motivation for the Application**

Travelling is not easy always, sometimes you get frustrated about getting the worst service at the airport, airlines or in airport lounges. Whenever you are planning for a trip or work you should know who is offering the best service for which you are paying which means *“Value for money”*. We are introducing our web application which lets you know the overall rating and reviews of different airports, airlines and airport lounges. We also have included the recommendation system for airlines based on similar ratings to provide the best offerings. Application is built upon the feedback received from thousands of users across the globe who have reviewed and rated the airlines, airport and airport lounges.

**Choice of Technologies:**

We have used different technologies for the different part of the project

**Apache Pig and Apache Hive -** For data pre-processing

* Data contained numerous null values and hence was required to pre-process the data
* Data cleaning was done in Pig and then it was loaded in hive, where we calculated the average ratings for airlines, airport and lounge.
* Then these different tables were exported as CSV to hdfs.

**Cloud SQL-** For database storage

* Created a database and tables in Cloud SQL.
* Loaded data from storage buckets to tables.
* Our Web App fetched data from this database.

**Google Cloud Platform(GCP) –** For application deployment

* We created a cloud bucket storage to store the final cleaned tables as CSV from hdfs
* Then we created a MySQL instance(cloud SQL) in GCP to create a database
* The web application was deployed in App Engine of GCP.

**Flask -** For web application development in python

* By referring to the Google Cloud Platform(GCP) documentation for developing a web app using the Flask framework in python.
* Flask played the role of a middleware where it fetched data from the database and displayed it on the web pages.
* We also used SQLAlchemy which is an open-source SQL toolkit and object-relational mapper for python programming. It was used for interacting with the database for defining models, fetching and filtering the data.
* The configuration files were updated to connect to the live instance on the GCP. To test on local machine a SQL proxy needs to be set up which will establish using a connecting link between the local application and database instance running on cloud.

**HTML/CSS -** For Web Page designing

* The Web UI designing was done in HTML
* Containers, tables, dropdown lists were added to create a web application.
* To apply styling to basic HTML tags, CSS classes were used.

**Jupyter Notebook –** For Implementing Machine Learning techniques

* Airline recommendation system was implemented using python pandas, numPy and matplotlib libraries in Jupyter notebook.
* Firstly, exploratory analysis on dataset was carried out for calculating the mean average ratings for the airlines.
* Graphs were plotted to determine the correlation between the overall ratings by a user for the airlines.
* Based on the correlation values, our model generates the top 10 airlines for the users based on the overall rating of the user.

**Data Source Selection:**

The dataset that we have used is a scraped dataset formed by scrapping all users reviews from the website Skytrax ([www.airlinequality.com](http://www.airlinequality.com/)). The website do not prohibit the scrapping and complete dataset is available on Github.

In total there are:

* 41396 Airline Reviews
* 17721 Airport Reviews
* 1258 Seat Reviews
* 2264 Lounge Reviews

**Pre-Processing:**

The data needed pre-processing before it can be used for our application. From the available reviews we have only used airline, airport and lounge reviews and we removed the unwanted columns from the other files. We calculated the average rating for every airline, airport and lounge from the ratings given by the user and stored it in different tables. We also calculated how many users have recommended a specific airline, airport or lounge.

Steps carried out for data pre-processing:

* Create a table with the necessary columns for airport, airlines and lounge.
* Load the date from original CSV file into the tables created.
* Run query to get average of ratings.
* Create a new table to store the average values.
* Save the file with average values into CSV file.

**Related Work**

In existing work, there are many websites that review the airlines and airports across the world. Airlines, airports and airport lounges are reviewed according to services offered by them to the users. In one of the research, authors have extracted keywords from user reviews and rating is given to each keywords using Hadoop and map-reduce and performed sentiment analysis to provide recommendations. In another research, work author built a recommendation system using Map-Reduce, Hive and Apache Mahout which is used to implement distributed and scalable machine learning algorithms.

**Challenges and lessons learned**

* In data processing, there were quite a few challenges faced as the previous assignment helped us get past it.
* Although with smooth workflow there were some glitches like the Date field datatype was object and we wanted that in DateTime format. We learnt how to convert it to the format we required.
* There was a problem in the data file where majority of cells were green tagged and was causing computational errors. We figured a way to clear the errors. Henceforth, no further computational errors reflected.
* This project helped us to understand the ETL(Extract, Transform and Load) operations on GCP.
* One of the biggest challenges was to figure out a way to export data from a Hadoop cluster to cloud storage bucket. After reading multiple GCP documentation we learned how it can be achieved.
* We also learned how the CSV stored in bucket storage is imported to a table of Cloud SQL database.
* One of the challenges was writing the configuration in the web application and also making the necessary changes to the GCP by enabling the APIs’ for connection.
* We also learned to develop an application using the Flask framework and SQLAlchemy toolkit.

**Responsibility statement**

Team Roles are as follows:

1. **Resource Investigator** – Anuja and Chethana searched and finalized the database and carried out pre-processing tasks on the database.
2. **Team worker** – Prateek and Omkar explored different databases and executed the connection part in GCP for MySQL. Vinit and Shivam explored the execution part of the code using Flask and Django frameworks in python and checked how to implement it in python.
3. **Co-ordinator** - Prateek and Vinit co-ordinated with the team to overcome the challenges.
4. **Plant** – Anuja and Chethana checked the inconsistencies in the database and gave the team insights on different aspects related to the data.
5. **Monitor Evaluator** – Omkar and Shivam were monitoring the changes that were being made by the team during the implementation of the code.
6. **Specialist** – Prateek and Omkar contributed in designing a recommendation system and wrote a code for implementing Machine Learning model in the project.
7. **Shaper** – Shivam and Vinit kept a check on the deployment part of the project on cloud.
8. Implementer – Omkar and Shivam worked on the database connection on GCP and Flask python code development for the system implementation.
9. **Finisher** – Prateek framed the report. Co-ordinated with the team for getting the final output files and created a demo video for presenting our team work and our project .