The Second Part of the Assignment of IDS 2019-2020

Introduction

The assignment guides you through the analysis of datasets using the techniques and tools provided in the course. This part of the assignment tests the understanding of the material in lectures 9-18. It is recommended to follow the assignment in the given order since the result of some questions might depend on answers to previous steps. The questions are detailed in the provided Jupyter notebook.

The Dataset

Dataset: In this part of the assignment the following datasets are used. For each question, use the data set as indicated in the description of the question.

- The provided data set "**air_pollution.csv**" contains hourly results of measuring the concentration of certain substances (CO, Benzene, NO2, particulate matter) and environmental conditions (temperature, relative humidity, traffic volume). This data was collected near a busy street in a city center by an automated device.
- The provided data set "**applications.csv**" contains information about the procedure of 13087 applications in an insurance company. Each row refers to an application and each cell shows the status of the application. For instance, the procedure for application in row number 4 was A_SUBMETTED, A_PARTLTYSUBMITTED and A_DECLINED.
- The provided data sets "pg_train.csv" and "pg_test.csv" contain a corpus of sentences from novels, labeled with the author (Austen, Chesterton or Shakespeare). Save for adding some spacing near the punctuation signs, this text has not been preprocessed.
- The provided data set "event_log.xes" collects events related to issues raised by clients to the help desk department of an Italian software company.

You should pass some steps before starting the assignment as preprocessing steps. The details of preprocessing steps are given in the Jupyter notebook file. After passing these preprocessing steps, export your final dataset as 'air_pollution_2.csv' dataset and use that for the

corresponding questions of the assignment. Make sure that you submit this extracted dataset with your results in Moodle.

Submission and Deliverables

The deadline for the assignment is 25/01/2020 23:59. You will need to hand in your submission via Moodle. Note that there is no extension for the deadline and also late submissions will not be considered. It is recommended to upload partial submissions, which can be replaced as you proceed with the assignment.

This part of the assignment should be done in the same groups as the first part. Make sure to include all group members' names and student ids in the submission!
Your submission should include a **Jupyter notebook**, which presents your results and also contains the python code used to obtain the results. Next to this Jupyter notebook, upload a zip-file that contains all requested data sets.

Report requirements:
You are allowed to upload 2 separate items via Moodle.

1. Jupyter notebook.
   - Use the provided Jupyter notebook to present results and code.
   - Make sure that the name and student id of **all the members** are in the Jupyter notebook.
2. datasets.zip including all the requested data sets.

Grading

Successful participation in the assignment, i.e. scoring at least 50% of the obtainable points, is one of the prerequisites for taking the written exam. The results of the assignment are valid for the current semester and expire afterward. The assignment can only be redone in the next academic year.

The grade of the assignment counts 40% towards the final grade. In this first part of the assignment, 100 points are obtainable, 90 points for the six main sections and 10 points related to your report style:

1. Preprocessing of the Data set – 5 points
2. Data Preprocessing and Data Quality – 10 points
3. Data Preprocessing and Advanced Visualization – 15 points
4. Clustering – 15 points
5. Frequent Item Sets and Association Rules – 15 points
6. Text Mining – 15 points
7. Process Mining – 15 points

- As a data scientist, adequately presenting your results is just as important as what you have done, therefore, 10 points are obtainable for report style.

Please note that the correctness of your code, its result and also the accuracy of your explanation are important.