Big Data Scientist & Engineer Individual Assignment 1

Course year: 2022-2023 Old program resits

Author: Evert-Jan Couperus

Version: 1.0
Date: 2022-09-01

Version control

Ver.	Status	Date	Author	Changes
1.0	Concept	2022-09-01	Evert-Jan Couperus	Changed individual assignment I from the Manual BDSE 21-22 into a separate document.

Table of Contents

1. Preface

Students that didn’t pass for the individual assignment I in the Big Data Scientist & Engineer course before course year 2022-2023 can still do the assignment as mandated by the previous Manual Big Data Scientist & Engineer version 2.2.
The assignment is the same as the official text in the manual. For the convenience of the student the assignment text is provided in a separate document. If this document contradicts the study manual, the study manual is leading.

2. Individual assignment I

This test involves all the skills / knowledge acquired in the first block of the semester. Learning goals for both the field of Big Data Scientist & Engineer are mentioned in detail in the weekly program.

Data requirements

The requirements for the datasets to be used are:

1. Kaggle: https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe
2. Your own scraping datasets of at least 10 labelled reviews
3. Your own hand-written dataset of at least 3 labelled reviews

Model requirements

A least 3 different classifiers should be built. Each of these classifiers should be capable of determining whether an additional hand-written hotel review is a positive or a negative review.

Part of the assessment is classifying at least 1 review provided by the teachers in max 10 minutes, so it is always a good idea to store your classifier model on a disk if it takes too long to run. Also, the predictions on the test set should be stored to gain performance during the assessment.

The mandatory deliverables are:

1. Python scripts, where
   - All the data is combined in one dataframe:
     - The Kaggle set
     - The webscraped set
     - Your own reviews
2. The total combined dataset is stored in a SQL database
   - The data used for Model building is fetched by a parametrized stored procedure
3. At least 3 types of classifiers are used to do a sentiment analysis
4. An extensive report, meeting the following requirements
   - In correct English (Dutch students are allowed to write the report in Dutch)
   - Containing only relevant screenshots of codes

- Clarify the process of
    - Data discovery
        - What datasets did you include?
    - Data preparation
        - How are the datasets stored?
        - What kind of processing was needed?
    - Model Building
        - Compare at least 3 different classifiers on overall accuracy on a test set
        - Explain the essence of the algorithm used by the classifiers
        - Overall performance
        - Possible fine tuning
- Communicate the results
- A checklist is added, see Chapter 7

And in addition

- The report should be uploaded to the DLO (no email attachments) at least 3 working days before the actual assessment date
    - No uploaded report ➔ no assessment
- Only the report is needed for uploading. No Python scripts!
- The report should meet the standards, i.e. there will be a checklist available for minimum requirements. If the report does not meet these minimum requirements ➔ no assessment

In our opinion these requirements cannot be met in less than 10 pages (including title page and index). This assignment is strictly individual.

3. Rubric Data Engineer and Data Scientist individual assignment I

Assessment Criteria – Data Engineer and Data Scientist individual assignment I		
Studentnumber:	Studentname:	Grading:

	Insufficient 0 - 25 points	Marginal 26 - 55 points	Good 55 - 75 points	Excellent 75 - 100 points
Data discovery	The student uses only the provided dataset and has little understanding of its content	The student has added a minimum of 10 hand written reviews and has turned the dataset into a data frame And the student has scraped TripAdvisor using the sample script.	Additional to *Marginal*: The student has scraped and labeled more than the minimum of 10 reviews from more than one hotel booking site and has turned the dataset into a data frame.	Additional to *Good*: The student has scraped and labeled more than the minimum of 100 reviews from several hotel booking site and has turned the dataset into a data frame.
Data preparation	The student can barely turn the provided dataset into a usable dataset of labeled data. There is no live connection with a SQL database.	The student can turn the dataset into a usable dataset of labeled data and perform some additional cleaning if needed. There is a live connection with a SQL database, no parametrized queries	Additional to *Marginal*: Moreover, parametrized querying is part of the script	Additional to *Good*: No embedded SQL is used in the script only stored procedures are used. More over some advanced cleaning had to be done
Model planning	The student has no idea about different models to be used for data science	Student only knows to describe the models involved in the script. But has no ideas about the pro's and the cons of the 3 models	Additional to *Marginal:* The student can explain the ranked accuracy of the 3 different models. In short, why is a model better than another?	Additional to *Good*: Student has done some research on classifiers, and can use arguments for using a particular one beyond the mandatory literature
Model building	Student cannot explain any of different statements in the code used to build a classifier. The dataset is not splitted into a training and a test set	Student can explain only the basic statements in the code behind only one classifier. The dataset is splitted into a training and a test set	Additional to *Marginal*: Student knows how to explain all the ins and outs of the pieces of code involved. In particular how to succeed in improving the overall accuracy	Additional to *Good*: Advanced tweaking of the parameters involved in the used classifiers has been used

Minimum requirement for a pass (i.e. grading ≥ 5.5):

- At least 3 out of 4 are *Good*
- None of them is *Insufficient*

4. Checklist Report

- □ Title page
- □ Table of contents (incl page numbering)
- □ Summary/abstract
- □ Introduction
- □ Background
    - □ Contains theory about the models
- □ Methods
    - □ Can contain multiple subsections
    - □ Screenshots of code, only when relevant
- □ Results
    - □ Contains relevant plots
- □ Conclusion and/or recommendations
- □ Reference list
    - □ Choose a consistent reference style: APA or IEEE

- □ Optional: preface, footnotes, appendices, list of symbols, glossary)
- □ Report is written in understandable and correct Dutch or English

Notes:
This checklist is used to check the completeness of the report, not whether the parts are accurate.

Only when your report is complete, you will be invited for the final assessment!

This checklist is derived from the ‘Beoordelingsformulier Onderzoeksrapport research skills/stage’.

If you need advice on how to write a report: tips can be found via the course ‘Research skills’ and online via the internship- and graduation manuals. (Accessible via DLO or A-Z).