Big Data Scientist and Engineer Individual Assignment 2

Course year: 2022-2023 Old program resits

Author: Evert-Jan Couperus
Version: 1.1
Date: Thursday, 02 February 2023

Version control

Ver.	Status	Date	Author	Changes
1.0	Concept	2022-09-01	E.J.T. Couperus	Separate document for assignment 2
1.1	Published	2023-01-30	E.J.T. Couperus	Explicit dataset requirement

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 II

Data requirements
The requirements for the datasets to be used are:

- Kaggle: https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

Assignment
Build

- a dashboard for interactive visualizations, which will give insight in for instance the datasets being used for applying the state-of-the-art-techniques
- a program applying DASK in the context of sentiment analysis
- a program in which at least 2 Neural Network techniques are applied. One could compare
  - Pytorch versus Keras
  - A regular Neural Network with a Convolutional Neural Network
  - A Convolutional Neural Network with a Recurrent Neural Network
  - ….

and write a report on your findings where

- overall performance and other quality measures are compared for both DASK and both the Neural Networks.

The mandatory deliverables are:

Python scripts, to be examined during an assessment

1. To obtain an attractive visual representation of all the data in the dataset, with visual interactive elements to support the so-called Visualisation mantra:
   - Overview: Gain an overview of the entire collection
   - Zoom: Zoom in on items of interest
   - Filter: filter out interesting items or filter in interesting items
   - Details: on demand: Select an item or group and get relevant information accordingly
2. All of the dataset is stored in a NOSQL database, for instance MONGODB. A live connection to filter data during the process of running the script should be implemented. Meaning during the process of filtering and plotting new data should be transferred from the database to the python script/ dashboard.
3. A script implementing DASK in the context of Sentiment Analysis
4. A script implementing at least 2 Neural Networks (of course it could be several scripts)
5. A comparison should be made on both scripts mentioned at 3 and 4 on performance and the overall accuracy / auc of both methods.

A compact report, to be uploaded on the DLO
The report should meet the following requirements

1. In correct English or Dutch
2. Containing relevant screenshots of codes
3. Containing relevant screenshots of the visualisation

4. Explain the inner working of DASK, Neural Networks in general and the specific Neural Networks you have implemented
5. Document your expectations on quality and performance
6. The checklist is applied

3. 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
  - □ Contain 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 'Reseach skills' and online via the internship- and graduation manuals. (Accessible via VLO or A-Z).

4. Rubric Assignment II

Insufficient if one of the 5 aspects described below is insufficient
Overall additional bouns of max 10 for excellent performance

	Insufficient	Marginal 10 points each	Good 12.5 points each	Excellent 15 points each
1.Visualisation	There is no dashboard, only one interactive plot inside the IDE	Only one of the minimum requirements is implemented (i.e. the mantra of Visualisation). There are no multiple tabs	All of the minimum requirements are implemented (i.e. the mantra of Visualisation). There are multiple tabs	Additional to *Good*: Sophisticated interactive elements are implemented
2. DB Storage NOSQL	There is no use of a NOSQL database	There is a live query for collecting data, however is a simple query, no aggregate is used.	There is a live query for collecting data, and the simple query is more or less complex, for instance an aggregate is used.	There is a live query for collecting data using the map reduce structure
3.DASK	There is no use of DASK	There is an implementation of DASK however the students has no ideas how it could be used	There is some research done after the way DASK could be used.	DASK is used with several machine learning algorithms and the student has successfully explored its capabilities
4.Neural Networks	There is no use of Neural Networks	Student only knows to describe the basics of the two neural networks involved in the script.	The student can explain both neural networks involved, the differences in the chosen approaches and some of the parameters involved.	Student has done some research on the topic, and is able to fine tune the model
5.Coding	Student cannot explain any of different statements in the code used to build a model.	Student can explain only the basic statements in the code behind the model	Additional to *Marginal*: Student knows how to explain all the ins and outs of the pieces of code involved.	Additional to *Good*: Advanced tweaking of the parameters involved in the used classifiers has been used.
6 Theory	Student has no basic understanding of ML	Student knows to explain a little of ML techniques and the role of Neural networks	Student knows exactly all ins and outs of a wide range of ML techniques	Student did research on his own, to explore in detail some ML techniques not covered by the course