Supplement Manual Big Data Scientist and Engineer Part 2

Author: Peter Odenhoven

Version: 1.5

Date: Monday, 12 April 2021

 

Course year: 2020-2021 special COVID-19 edition

Version control

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ver.** | **Status** | **Date** | **Author** | **Changes** |
| 1.0 | concept | 2018-10-01 | P.Odenhoven | Split the manual in Part 1 and Part 2 |
|  |  | 2018-10-04 | Z.Efendieva | Mimimize assignment  Choice between   * Either parallel processing with Spark * Or Neural networks |
|  |  | 2018-10-06 | S.Robben | Minimize program, editorial comments |
| 1.3 |  |  | P.Odenhoven | Review |
| 1.4 |  | 2020-11-04 | P.Odenhoven | Covid version, weekly program removed to external \*.xls |
| 1.5 |  | 2020-11-04 | Z. Efendieva | Updated: name of the subject, small corrections |
| 1.6 |  | 2021-04-09 | P.Odenhoven | Updated Assignment 2 ( Due to Covid) |

**Table of Contents**

[1. Examination and tests in block 2, change because of COVID-19. 4](#_Toc69125959)

[2. Test 3: Individual Assignment II ( Corona restrictions combines the assignment and the written exam) 5](#_Toc69125960)

[3. Test 3: Checklist Report 7](#_Toc69125961)

[4. Weekly planning Part 2 8](#_Toc69125962)

# Examination and tests in block 2, change because of COVID-19.

Important weeks for the examination/ tests:

|  |  |
| --- | --- |
| Week 11-20, Block 2 | Description tests |
| Week 13 | Resit Assessment Individual assignment I ( = test 1) |
| Week 18-20 | Assessment Individual assignment II ( = test 3) |
| Week 19-20 | Resit Assessment Individual assignment II ( = test 3) + Mathematics test (= test 2) |

Because of Covid-19 there will be no written exam ( = schriftelijke toets). In SIS your grading for test 3 is registered for both the individual assignment as for the so-called combi-toets.

# Test 3: Individual Assignment II ( Corona restrictions combines the assignment and the written exam)

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

1. 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.
2. A script implementing DASK in the context of Sentiment Analysis
3. A script implementing at least 2 Neural Networks ( of course it could be several scripts)
4. 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

# Test 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).

# Weekly planning Part 2

See the uploaded \*.xls on the DLO.