Manual Data Engineer and Data Scientist Part 1

Course year : 18-19

Author: Peter Odenhoven
Version: 1.4
Date: Friday, 06 September 2019

Version control

Ver.	Status	Date	Author	Changes
1.0	Concept	2018-06-11	P.Odenhoven	• Rubric Individual assignment I • Rubric Individual assignment II • Theoretical Exam
1.1	Semi Definitive 18/19 sem 1	2018-8-20	Z.Efendijeva S.Robben	• Improved Rubric • Textual marks for learning goals not covered by datacamp • Not yet final: Rubric individual assignment II
1.2		2018-10-01	P.Odenhoven	• Split into Part1 and Part2
1.3		2019-01-10	P.Odenhoven	• Review DC courses • Review Individual Assigmnent I
1.4		2019-08-30	Z. Efendijeva	• Changes in planning of Mathematics / test

Table of Contents

1. Preface

The semester on Big Data has been operational since jan 2016. At the time there were 2 modules Data Processing and Data Mining & Analysis, each for 4 European Credits. The sequence for the modules has changed several times, in order to support the project work more to its demanding. However, all the projects on Big data are characterized with a slow start: it takes time for students to get oriented on the dataset and the problem involved. Moreover, in at the end of most the projects time was lacking for a thorough analysis of the data since the theory involved was not yet taught during classes. So, these are the reasons why we have decided to make the semester more a-symmetrical, meaning in the first part of the semester, the emphasis will be on getting to know all the algorithms and aspects of data engineering and data science the theory. Whereas the second part of the semester the emphasis will be on the project work. As a consequence, the former modules are merged in to a new one called Data Engineer and Data Scientist, making it possible to

- React more flexible on what is needed for the project
- To shift lessons and workshops from the second part of the semester to the first part

2. Introduction

The module Data Engineer and Data Scientist is part of the thematic semester Big Data. During this semester you will be confronted with different parts of the so-called Data Science Life Cycle, see figure below [1]

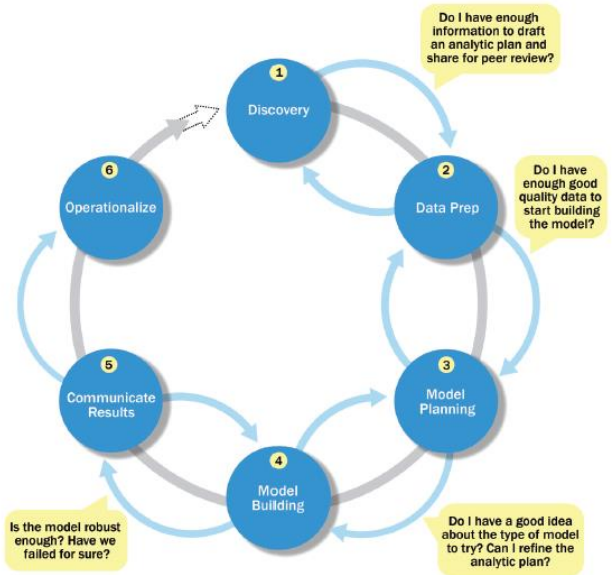

FIGURE 2-2 *Overview of Data Analytics Lifecycle*

So, the following learning outcomes can be formulated

1. The student searches, selects and collects different types of data; being structured and/or unstructured.
2. The student is familiar with all kinds of different storages, be it SQL or NOSQL
3. The student analyses the business data and develops datasets, that have an optimal fit with the organization's information needs.
4. The student applies the aspects of the data comprehensive, like for example consistency, granularity, normalization and the range of values.
5. The student explores, pre-processes and conditions these data and uses them for advanced analysis and modelling.

[1] *Data Science and Big Data Analytics, Discovering, Analysing, Visualizing and Presenting Data (EMC educational services)*

6. The student understands the mathematics involved in most of the data mining algorithms
7. The students can make motivated choices regarding the use of several Data Scientist algorithms and regarding the specific parameters within an algorithm
8. The student can make motivated choices regarding the use of tooling for Data Engineering and Data Scientist
9. The student can make an analysis of a rather large dataset using the tool which suits you and the problem the best

This manual has the following structure:

- Chapter 2 describes the overall organisation of the module
- Chapter 3 describes the organization of the exams
- Chapter 4 is an overview of all 4 exam components
- Chapter 5 describes the requirements for the individual assignment I (= test 1)
- Chapter 6 contains the rubric for the individual assignment I (= test 1)
- Chapter 7 contains a checklist for the Report on Machine Learning (=test 1)
- Chapter 8 contains a rubric for the theoretical exam (= test 4)
- Chapter 9 describes the weekly program. All the mentioned learning goals in the weekly program are mandatory for the theoretical exam
- Chapter 10 lists most of the needed software/ hardware

Dhr. P. Odenhoven

3. Overall organisation of Data Engineer and Data Scientist

During the semester you get to know all kinds of topics related to the field of Data Engineering and Data Scientist. From the website[2] you can learn that

Data engineers build massive reservoirs for big data. They develop, construct, test and maintain architectures such as databases and large-scale data processing systems. Once continuous pipelines are installed to—and from—these huge "pools" of filtered information, data scientists can pull relevant data sets for their analyses.

Data scientists are big data wranglers. They take an enormous mass of messy data points (unstructured and structured) and use their formidable skills in math, statistics and programming to clean, massage and organize them. Then they apply all their analytic powers—industry knowledge, contextual understanding, skepticism of existing assumptions—to uncover hidden solutions to business challenges.

So, the module is built out of 3 components,

1. Data Engineer

- The basics of SQL (in case you are missing these basic skills) and a little more
- The basics of relational Databases
- Programming: R or Python. More technical interested students are allowed to learn Python, but they should be aware of the fact that
    - Interactive visualisations are "easier" to program in RShiny than in plain JavaScript
    - Visualisations in for example D3 JavaScript are known for bad performance when it comes to big datasets
- The basics of NOSQL databases
- The basics of HADOOP
- The basics of Spark

Skills in the field of Data Engineering are tested by individual assignments and a theoretical exam at the end of the semester.

2. Data Scientist

- Classic data mining involving techniques like
    - Classification/ prediction
    - Regression

[2] https://medium.com/@vegi/data-scientist-vs-data-analyst-vs-data-engineer-using-word-cloud-902ab83d0879

    - Clustering
    - Text Mining
    - Recommendation
- Techniques involving among others
    - Spark
    - Neural Networks
    - Image recognition
- Students need to understand the algorithms involved. They need to know why a specific algorithm is preferred within a certain context and which parameters can be used to influence the outcome of the algorithm. The algorithms are implemented in most data science libraries

Skills in the field of Data Scientist are tested by individual assignments and a theoretical exam at the end of the semester.

3. Mathematics

Fundamental for most machine learning algorithms for is a (small) notion of basic concept of Mathematics:

- Distance measures
- Probability and descriptive statistics
- Comparison of sets
- Correlation and regression
- Basics of Linear Algebra

Skills in the field of Data Scientist are tested by online (MapleTA) test. To be able to participate in this test, student has to complete a homework first (available from week 5 in MapleTA).

Each semester at the HvA consists of 2 blocks, i.e. a period of 10-12 weeks. Since in our experience, Big Data projects start off slowly, students first need to understand the problem, the data structures involved, the required tooling etc., and on the other hand students need to learn the theoretical background of the machine learning algorithms that is why we have chosen to have an emphasis on theory more in the beginning of the semester and applying in during the project at the second part of the semester:

Block 1	Lessons/ hours of contact	Block 2	Lessons /hours of contact
Data Engineer / Data Scientist	3 * 2	Data Engineer / Data Scientist	1 * 2
Mathematics	1 * 1		
Project coaching	1 * 3	Project coaching	1 * 7

Mandatory E-learning environments

1. For both the lessons on Data Engineering and Data Scientist the skills to learn rely heavily on the DataCamp website courses https://www.DataCamp.com/home
    - Either Track: Data Scientist with R
    - Or Track: Data Scientist with Python

An Academic License is available and will be supplied to you whenever it is needed. The courses combine acquiring practical skills in R or Python and easy to understand videos for explaining the theoretical backgrounds are explained.

2. For getting to know the technical implications of Big Data the Cognitive Class website courses https://cognitiveclass.ai/ is used among others.
3. Last but not least the course on Mathematics will be taught using a E-learning environment (T-Maple) where students can practice their skills and get immediate feedback on the answers. Students who followed the Essential Skill course in their freshmen year, should be familiar with this environment.

About all the topics on Data Engineering and Data Scientist are supported bij DataCamp videos or other online sources. For some topics additional slides will be provided. Students who prefer to study theory from a book rather than a bunch of websites are advised to buy the book:

Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data. EMC educational services

The book is one the most readable books on the market, when it comes to the balance in elaborate mathematics, coding and theory.

4. Examination and tests

This chapter describes the examination and grading for the course Data Engineer and Data Scientist.
The course is rewarded with 8 European Credits (studiepunten) if and only if the students has passed 4 test, with 4 separate grades:

			ec
Block 1			
1.	Data Engineer and Data Scientist:	Test 1: Individual assignment I + Report	4
2.	Mathematics:	Test 2: MapleTA test, graded >= 5.5	1
			5
Block 2			
3.	Data Engineer and Data Scientist:	Test 3: Individual assignment II + report,	1
4.	Data Engineer and Data Scientist:	Test 4: Overall Theoretical exam	2
			3

Important notes:

- All 4 separate grades for the tests should be sufficient, i.e. ≥ 5.5
- The overall theoretical exam will cover all the topics from Data Engineer and Data Scientist, including knowledge on SQL and NOSQL database, Hadoop, Map Reduce and Spark. In the weekly program all learning goals are explicitly listed. This list should be your checklist, when preparing for the theoretical exam.
- The individual assignments are individually assessed in a 10 minutes window at the end of each block. Students are only allowed to attend the assessment after the report is uploaded.
- Participation in Mathematics test is only available if students have finished their homework in MapleTA (available form week 5).

There is only one bonus opportunity for test 1:

Block 1

1. There will be an intermediate test on your R/ Python skills. This test can be rewarded with a maximum of 2 bonus point
2. There is no resit for the bonus test. If you miss the test in week 4, then it is too bad, but the consequence is no bonus!
3. The bonus will be added directly to the final grading of your individual assignment I (= test 1), with the following restrictions
    a. Grading for the final assignment cannot exceed a 10-score
    b. The earned bonus in the intermediate test, will not be valid for any resit. If you have to resit the assignment the bonus will be expired. It is only valid if and only if you take the individual assessment the first time it is planned (week 8,9 or 10 of the semester)

Important weeks for the examination/ tests

Week 1-10, Block 1	Description tests
week 4	Intermediate bonus test for assignment I (= test 1) (no resit)
Week 5	Math homework is available (closes a day before the test)
week 7	Math test
week 9-10	Assessment Individual assignment I (= test 1)
Week 11-20, Block 2	Description tests
week 18-20	Assessment Individual assignment I (= test 3)
week 18-20	Theoretical exam (= test 4)
week 19-20	Resit Assessment Individual assignment I and II (=test 1 + test 3) (bonus points are invalid!)
week 19-20	Resit Math test

5. Test 1: 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 Data Engineering and Data Scientist 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 a 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 15 minutes, so maybe it is a good idea to store your classifier model on 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. R scripts or 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
    - Overall performance
    - Possible fine tuning
- Communicate the results
- A checklist is added, see Chapter 7

And in addition

- The report should be uploaded tot the VLO (no email attachments) at least 3 working days before the actual assessment date
  - No uploaded report ➔ no assessment
- 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
- Only After uploading a report which satisfies the minimum requirements the student will receive a time slot for the assessment

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

6. Test 1 : 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 uses only the dataset provided and has appropriate insight in its content. The student has scraped only the minimum of 10 reviews from a hotel booking site and added the minimum of its own written reviews	Additional to *Marginal*: The student has scraped at least 100 reviews from a hotel booking site and has turned the dataset into a data frame. Moreover, the script be used for live scraping upon request. The student is able to add written reviews on request.	Additional to *Good*: The student has scraped more than 100 reviews from more than on hotel booking site.
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 connection with a SQL database, no parametrized queries	Additional to *Marginal*: The student can turn the *combined* set into a usable combined dataset of labeled data and perform some additional cleaning if needed. 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. More over the student can explain the predictions	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* and None of them is *Insufficient*

7. Test 1: Checklist Machine Learning 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 VLO or A-Z).

8. Test 4: Theoretical exam

The theoretical exam involves all theory and obtained skills for the module Data Engineer and Data Scientist. It includes also the theory of the Mathematics lessons.

Learning objectives	Type vragen					
	Reproduction	Production				
	Understanding Remembering	Applying	Analysing	Evaluating	Creating	
1. Understand the mathematics involved in most of the data scientist algorithms	10%	10%				20 %
2. Make motivated choices regarding the specific parameters within an algorithm		10%	10%			20 %
3. Understand the outcomes of an algorithm used to analyse a dataset		10%	10%			20 %
4. Make motivated choices regarding the use of several Data Scientist algorithms			10%	10%		20 %
5. Advise on storage issues				10%		10 %
6. Advise on a Data Engineering and Data Scientist stack					10%	10 %
Total	10 %	30%	30 %	20 %	10 %	100 %

Moreover

- All questions will be open questions, no multiple choice.
- Questions are posed in English, students are allowed to answer in Dutch
- No calculations need to be done, so no calculators are needed
- Learning objective 5-7 are tested with a case study

9. Weekly planning

Each lesson has an indication on the topic

- DE = Data Engineer, all learning goals are mandatory for the theoretical exam and some of them apply to the individual assignment
- DS = Data Scientist, all learning goals are mandatory for the theoretical exam and most of them apply to the individual assignment
- Workshop, handy hands-on practical for either the individual assignment or the project task

The course is supported by material on the E- learning of the HvA. Most important will be the folder:

Big Data semester/Documenten/03 DATA ENGINEER AND DATA SCIENTIST/02 DATA ENGINEER AND DATA SCIENTIST Weekly Material students where you will find additional slides, exercises, datasets, need to knows etc. Learning goals not covered by slides on DataCamp are marked with an asterix (*) and should be covered by additional slides.

Lesson	BLOCK 1		Learning goals
	Week 1	remarks	
1 **lecture**	Introduction on Big Data Cognitive Class https://cognitiveclass.ai/courses/ • Big Data 101	Create Cognitive Class account Subscribe to invitation mail from DataCamp Start Installing software, see list Appendix	• Gain insights on how to run better businesses and provide better services to customers by machine learning algorithms • Understand to process big data on platforms that can handle the volume, velocity, and variety (3 V's) of Big Data --- • Know the Pros and Cons of R and Python
2 **workshop**	First steps with a data analyst tool	Hands-on exercise: Rstudio Intro IDE Python tutorial	• Get to use RStudio / Python IDE ○ Get working directory ○ Set working directory ○ Packages / libraries ○ R / Python version ○ Making a project Running a script

3 **workshop**	First steps with R DataCamp R • Introduction To R DataCamp Python • Intro to Python for Data Science	Hands-on exercise: R Basic operations Python Pandas Dataframe exercise Python Plot exercise	• Get to know the basic datatypes ○ Scalars ○ Vectors ○ Matrices ○ Dataframes ○ Lists ○ … Know to explain • Why do we need variables? • Why do we need functions? • How to implement conditional flow: if … then … • How to loop: for • How to build a function • How to avoid looping with apply functions in R
4 **Mathematics**	Basic set operations Basic statistics		Know how to calculate and interpret • Mean value • Five number summary • Standard deviation • Cartesian product
Assignment I	Getting the Kaggle data Explore other useful websites	Kaggle dataset	

	Week 2		
1 workshop	Different types of datasets DataCamp R • Importing Data in R (Part 1), *.csv,*.xls,*,xlsx DataCamp Python • Importing Data in Python (Part 1)	Hands-on exercise R: Cleaning data Python cleaning data	Knowing • How to import flat files or other formats • How to write to a flat file • How to make a connection with a local SQL database • How to extract data from a database • How to write data to a database
2 workshop	Manipulating data in R DataCamp R • Data manipulation in R with dplyr DataCamp Python • Panda Foundation	Hands on exercise R: dplyr Python: dfply	Knowing • How to recognize messy data • How to deal with messy data • How to combine data in different dataframes • How to aggregate data in a dataframe
3 lecture/ workshop	Old skool SQL DataCamp R • Intro to SQL for Data Science DataCamp Python • Introduction to Database in Python	Query on a local MySQL database Python database exercise	Know to explain • Why do we need databases? • Why do we need relational databases? • What is are downsides of relational databases? • How to query a database with SQL ○ Where … ○ Aggregation, group by ○ Inner Join, Left join, Right join, Full Outer join ○ Union / Minus (Except) * ○ Parameterized queries * • Why do we need parameterized queries? * • Why do we need stored procedures? * • How to connect to a local database? * See 02 SQL1819.ppt

4 Mathematics	Probability		Knowing how to calculate • Conditional probability • Independency of probabilities
Assignment	First insight in data (quality) by • Storing it in a SQL database • SQL queries		

	Week 3		

1 lecture	Introduction Machine Learning
	
	DataCamp R
	• Introduction to Machine learning
	
	DataCamp Python
	• Supervised learning with scikit-learn
		Demo Women having an affair	Knowing the answers for
			• What is Unsupervised learning?
			• What is Supervised learning?
			  ○ What is Regression?
			  ○ What is Classification?
			  ▪ Which different types are around?
			  ▪ What is a Confusion matrix?
			  • Calculate accuracy
			  • Calculate TP ratio/ FP ratio
			  • Calculate TN ratio/ FN ratio
			• What is Cross Validation
			  ▪ How to prevent Overfitting
			• What is meant by a ROC curve
			• When do we need Unsupervised learning?

2 workshop	Classification
	
	DataCamp R
	• Supervised learning with R
	
	DataCamp Python
	• Linear classifiers in Python
		Hands on exercise Random Forest	Identifying different classification problems
			
			Knowing and understanding
			• The pros and cons of
			  ○ Naïve Bayes
			  ○ Decision tree
			  ○ Random Forest
			  ○ K Nearest NeighBour

			○ Logistic regression ○ Support Vector Machines • The fundamentals of the algorithms involved • The influence of the parameters of the algorithms involved
3 workshop	Correlation and Regression DataCamp R • Correlation and Regression DataCamp Python https://pythonspot.com/linear-regression/	Hands on exercise Regression	Knowing • How to calculate correlation • The difference between correlation and causation • How to interpret correlation • How to interpret the linear model • The Regression output terminology, among others RMSE • How to interpret the regression coefficients • How to interpret R squared
4 Mathematics	Regression		Understanding how to calculate • Line through points • Regression line • Spearman rank correlation
Assignment	• Store it in a SQL database • Trying to figure out whether it suits your needs		

	Week 4		
1 **Bonus test**	Intermediate bonus test		Note: For the bonus test there are no resits
2 **workshop**	Classification DataCamp R • Support vector machines DataCamp Python Linear classifiers in Python	Hands on exercise Logistic Regression and SVM Textmining	Identifying different classification problems Knowing and understanding • The pros and cons of ○ Naïve Bayes ○ Logistic regression ○ Decision tree ○ Random Forest ○ Support Vector Machines • The fundamentals of the algorithms involved • The influence of the parameters of the algorithms involved
3 **lecture**	Text mining / Sentiment analysis DataCamp R • A bag of words DataCamp Python • Natural Language Processing Fundamentals in Python	Demo	Knowing • How to Tokenize • N- gram • Why mostly a Corpus is built • How to clean a Corpus • What is meant with a Document term matrix or Term document matrix • What is meant with sparsity in this context • The goal of TF_IDF • How to calculate the TF_IDF in a certain case • The pros and cons of • Naïve Bayes
4 **Mathematics**	• Distance & similarity		Knowing how to calculate different distance measures • Euclidean distance • Manhattan distance • Cosine & Jaccard similarity Knowing Pros and cons of different distance measures

Assignment	• Cleaning the data • Making training sets and test sets • First try / first classifier		

	Week 5		
1 workshop	Sentiment Analysis	Hands on exercise	
2 lecture	Unsupervised learning DataCamp R • Cluster analysis in R DataCamp Python • Unsupervised learning in Python	Demo	Knowing and understanding • How to estimate the number of clusters • How to evaluate a cluster • The difference between hierarchical clustering and k- means clustering • Why PCA is sometimes needed • How to evaluate the PCA features
3 workshop	Webscraping DataCamp R • Working with Web Data in R DataCamp Python • Importing Data in Python (Part 2)	Hands on exercise	Understanding Application Programming Interface • Using API clients Knowing • The difference between a GET and POST request * • How to manipulate JSON • How to manipulate XML • How to get around rate limiting * http://toolsqa.com/postman-tutorial/
4 Mathematics	Basics of Linear Algebra		Knowing how to • multiply a matrix with a matrix • multiply a matrix with a vector • eigenvalue and eigenvector
Assignment	• Scraping the web, building your own dataset • Creating basic plots • Interpreting confusion matrices • Set up report		

	Week 6		

1 lecture	Recommender
	
	R
	• http://rstudio-pubs-static.s3.amazonaws.com/248530_18970dc8eb4046a6b4f2fba987fe2a50.html
	
	DataCamp Python
	• Tutorial: Recommender Systems in Python Beginner
		Hands on exercise	Knowing the answers to questions like
			• What is collaborative filtering? *
			  • User based
			  • Item based
			• What is a content-based recommender? *
			• What are the pro's and cons of the different kind of recommenders? *
			
			* 01 Recommendation 1819.pptx

2 3 consultation	Teachers are available for help with the individual assignment		
4 Mathematics	Homework		• No lecture. Finish homework in MapleTA, before the test.
Assignment	• Pushing the classifiers to obtain higher accuracy • Finishing Report		

	Week 7		
1 2 3 consultation	Teachers are available for help with the individual assignment		
28 Mathematics	MapleTA test (requirement: homework is complete)		
	Week 8-10		
	Assessment of Individual assignment I (= test 1)		

10. Checklist Report for the individual assignments I and II

- ☐ 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).

11. List of needed Software (not at all trying to be complete …)

Name	Type software	Mandatory	requires	url	remark
1. JAVA version 1.8.0	JRE	YES		https://www.java.com/nl/download/	Java Virtual Machine
2. MySQL	DataBase	YES	Oracle account	https://dev.mysql.com/downloads/	Community download. On this page you can also find connectors and the workbench
3. MySQL shell	shell	YES			
Either	MySQL workbench	optional		https://dev.mysql.com/downloads/	Sometimes it crashes
Or	HeidiSQL	optional		https://www.heidisql.com/	Nice light weighted frontend
4. R3.3		YES, for R class	For most packages	https://www.freestatistics.org/cran/	RStudio can R different R version
5. R3.4		YES, for R class	For some packages	https://www.freestatistics.org/cran/	
6. RStudio	IDE for R	YES, for R class		https://www.rstudio.com/products/rstudio/download/	
7. Anaconda (Python 3)	Data science platform	optional		https://www.anaconda.com/download/ https://stackoverflow.com/questions/34097988/how-do-i-install-keras-and-theano-in-anaconda-python-on-windows	Jupyter notebooks
8. Python IDE	IDE for Python	YES, for Pyhton class		Check out : https://www.datacamp.com/community/tutorials/data-science-python-ide	

9. NO SQL Databases					
10. MongoDB	Database	YES		Check out : https://www.mongodb.com/download-center#community	
11. Studio3T	MongoDB shell	YES		Check out: https://studio3t.com/	