# Individual assignment I , course year 2020-2021

This assessment/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 build. Each of these classifiers should be capable of determining whether an additional hand-written hotel review is a positive or a negative review. 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. 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, at least
     + A wordcloud
     + A plot given a first insight in the dataset, for example
       - Number of reviews for each hotel
       - Number of reviews for each nationality
       - …
   * 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
       - Explaining in your own words for at least 3 different classifiers how the algorithm obtains the right classifications
       - Compare the 3 different classifiers on overall accuracy and other appropriate measures on a significant test set
       - Possible fine tuning , at least one technique shoud be used to try to improve the performance of at least one classifier
   * Communicate the results
   * A checklist is added

And in addition

* + The report should be uploaded tot the DLO ( no email attachments) at least 2 working days prior 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.

# Checklist Machine Learning report assignment I

* 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 ‘Reseach skills’ and online via the internship- and graduation manuals. (Accessible via VLO or A-Z).