# Hadoop Assignment 1: MrJob

Name: Chris van Roode

Studentnumber: 627687

Github: https://github.com/CBvR/Hadoop\_Assignments

## Setup

Have an Hadoop virtual machine installed with MRJob, PIP and the delivered u.data file (which should be in the same folder as the code file

## Execution

Be sure to be logged in as root

To run the python script:

* On the VM:

python Chris\_vanRoode\_Hadoop\_Assignment1.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoopmapreduce-client/hadoop-streaming.jar u.data

* Local:

python Chris\_vanRoode\_Hadoop\_Assignment1.py u.data

## About

The goal for this code is to get all the movie id’s according with their ratings. To do this, i started with fetching all movie id’s and the list of their ratings. After that i grouped the ratings so every movie id has 1 record showing. This then needed to be sorted starting from the highest rating.

To get all the movie id’s i used the mapper function, to get all the movie\_id’s and their ratings linked to the id’s.

To sum the ratings of each movie, i combined all the ratings of each individual id with the combine function.

At last, the sorter is used to see all the ratings from the movies sorted starting from the highest value.

## Github

Code file: Chris\_vanRoode\_Hadoop\_Assignment1.py

Document: Chris\_vanRoode\_Hadoop\_Assignment1\_documentation.docx

Result of code: Capture.png

## Result

Afbeelding met tekst, map, elektronica

Automatisch gegenereerde beschrijving