Data Scientist at 42matters
Interview Task

General Requirements:

We are looking for a good code structure and clear communication style. Commands in the READMEs must work out of the box and code should not contain passwords, credentials and hardcoded paths. Python as a language has naming conventions that should be followed. Documentation should be easily readable, Markdown (what Github uses) is a good way to do that. Code should be commented. Tools and commands used should contain a very clear instruction on how to install and configure. A good and platform-independent way is to do this via Docker, but we leave that open to the candidates in case they have a better way to do that.

1. **Movie categorization**
   You are given millions of movies and a list of thousands of movie categories (names only e.g. “Sci Fi Movies”, “Romantic Movies”). Your task is to assign each movie to at least one of the movie categories. Each movie has a title, description and poster.
   - How would you solve this problem?
   - How would you verify its quality?
   - How would you handle the case of adding or removing a category?
   - How would you handle the case of adding or removing a movie?

   DELIVERABLES:
   - Description about the chosen approach and its pros and cons (no code required)
   - Short discussion about alternative approaches you might have considered and their pro and cons

2. **Word count in PySpark**
   The goal of this task is to count the words of a given dataset.

   The tasks to do are:

   a. Download the data set over which to run word count from the following link: https://s3.amazonaws.com/products-42matters/test/biographies.list.gz

   b. Implement a PySpark program that counts the number of occurrences of each word in the provided file. Only lines starting with the “BG:” should be considered,

and a whitespace tokenizer should be used for tokenizing the text.

DELIVERABLES:

- Code
- Documentation that explains how to run the code
- The result of the word count

3. **Movies view estimations in Python**

The goal of this task is to implement a model in python to estimate the number of views a movie has.

You have the following data available:

- Movies list of 250 movies: https://github.com/WittmannF/imdb-tv-ratings/blob/master/top-250-movie-ratings.csv which contains the position of 250 movies as well as their rating count and other information
- A limited number of movie views data
    - Forrest Gump: 10000000 views
    - The Usual Suspects: 7500000 views
    - Rear Window: 6000000 views
    - North by Northwest: 4000000 views
    - The Secret in Their Eyes: 3000000 views
    - Spotlight: 1000000 views

DELIVERABLES:

- Code
- Documentation that explains how to run the code
- Description about the chosen approach and its pros and cons
- Short discussion about alternative approaches you might have considered and their pro and cons