**Data Science Test Case**

**Intro** 

Hi!   
Thank you very much for your participation in our recruitment process for the Data Science role! 

After looking at your CV and speaking with you we strongly believe you are a good Candidate, and the next step is to confirm your capabilities with this hands-on challenge.  

The following test has three parts:

* The **first** part will be verifying your NLP-specific Data Science expertise. Here we will put weight on:
  + How you plan, structure, and prepare your experiments
  + How you implement your experiments
* The **second** part is included to test your knowledge regarding algorithms and data structures. here we will put weight on:
  + What thoughts you have made around the performance of your implementation
  + How you implement your search algorithm
* The **third** part is included in order to hear about your experience with all the stuff around the algorithms. Here we will put weight on:
  + Your knowledge of MLOps principles
  + Your understanding of system architecture

We estimate that these three parts should consume 6-7 hours of work. Hence, we give you 5 working days (7 calendar days) and we put the deadline after we confirmed you received the email from us. In case of personal time constraints, you are not able to deliver your results – please let us know beforehand – we can agree on a more suitable timeframe for you.  

It is important that you present your code and conclusions in a structured and easy to read manner.  If you have any questions, please do not hesitate to contact us!  

We appreciate that these coding parts might be a challenge to accomplish in a way that fulfills the requirements 100%. If you do not manage to find solutions for both the tasks – please do not give up and send what you have done. 

Good Luck and have fun!

**PART 1 – Data Science expertise**

You should have received a data file called “data science part dataset.7z”.

Data Description – “data science part dataset.7z”:

The dataset contains the two folders: *testing\_data* and *training\_data*. Both folders have identical structure. Each folder contains the two subfolders: *annotations* and *images*. In the folder *images* you will find a series of .png files. In the *annotations* folder you will find a series of .json files. The .json files contain data as it would have been returned from a Optical Character Recognition (OCR) service had the images ben sent to the service. Each .json file has a corresponding image in the *images* folder. The .json and image files are linked by having the same name. The .json files contain the following variables for each text found (a text contains one or more words):

* **text:**  The text containing one or many words bound together
* **label:** The categorization that the OCR service has given the text
* **words:** The words or tokens contained in the text
* **id:** The identifier of the text
* **linking:** The ids of the texts that the specific text is linked to
* **box**: Traditionally, in OCR, the axes of the bounding boxed start at the top left corner and go right (X) and down (Y), so coordinates represent this:



You are allowed to make certain assumptions to narrow down the scope of your work. This includes not using all the data given. Please share the assumptions in a readme file.

**For this task it is preferred (but not necessary) that you use python.**

Using the data you are asked to plan and conduct one or more data science experiments.

An experiment is defined as:

1. Using the training\_data to train or fine-tune an algorithm to classify the **text** variable.   
   You can use any additional variable or variablesas input data. Use the **label** variable as the target value
2. Evaluate the model using the testing\_data.  
   Visualize, report, and attempt to explain the findings.

**PART 2 – Algorithms and Data Structures expertise**

You should have received a data file called “algorithms part dataset.7z”

Data Description – “algorithms part dataset.7z”:

The dataset contains one .csv file. The file contains data from the video game vendor Steam. The .csv file contains the following five variables:

* **ID:** Theunique identifier of the
* **GAME\_NAME**: Name of the game
* **BEHAVIOUR:** Behavior name (purchase/play)
* **PLAY\_PURCHASE**: Hours if behavior is play, 1.0 if behavior is purchase
* **NONE:** Not in use

You are allowed to make certain assumptions to narrow down the scope of your work. This includes not using all the data given. Please share the assumptions in a readme file.

**We prefer that you use other language than Python for this task**. Additionally, do not use any libraries with deduplication or string analysis methods.

Using the data, you are asked to:

1. Construct a tree-based data structure. It is up to you to decide in what order the tree will branch.
2. Build a search algorithm to traverse the tree and spot the duplicates (in the .csv file this would be two rows with same values for all the columns). Mark the nodes and print to command prompt or similar.

**PART 3 – Machine Learning Operations**

There is no coding done for this part.

Instead, you are asked to think about the following product.

Product Description:

You are working on a project to create a platform that reads and extracts information from millions of invoices a year. The platform automates the process of invoice creation and payment for customers. The platform will use ML models to predict and extract data from the many different types and layouts of invoices based on past data.

During the interview you will be asked how you imagine an MLOps architecture capable of supporting such a product could look like. This includes the technology stack and development flow.

Think about subjects such as:

* Data Collection & Management
* Model Development
* Model Deployment & Management
* Model Monitoring & Maintenance
* Application Development and integration