LOC·ANALYTICS

Madrid Studio

Associate Data Scientist

Technical Test

2020-07-07

INTRODUCTION

We need to know your technical skills of Data Scientist through a practical test. The test does not exclude selection processes but allows us to know your skills.

The test shall assess:

1 **BIG DATA ENGINEER**

- Skills of programming, Databases, Data Quality, Basic Statistics, Machine Learning, NLP, etc.

2 **TECHNICAL**

- Technological skills are fundamental: programming, Algorithms& Data Structures, etc

3 **BUSINESS AND STORYTELLING**

- Know your communication skills and business value: Strategy, Value propositions on the case, Plan of action for Implementation

You can sophisticate the response of the case as much as you want

TECHNICAL TEST

We start with the technical test, consider the following milestones…

Data Science	• The test will be evaluated by the EA Loc Analytics technical team
Informed of the case	• Description of the objective and data for the technical test are provided.
Development of the Test	• Resolution of the case. • It is recommended to use Open Source tools or free platforms. • The evaluation of the test will require the delivery of the code (preferably in a **Github repository)** • You can add any Open Data information you deem appropriate
Storytelling	• The results will be presented in writing in a repository (preferably Github) and in person to the EA team. Present to us the results obtained in the way you consider convenient. We are open to all options

NOTE: In the case of using tools other than Open Source or Data Sources not free for the case, Electronic Arts (EA) is not responsible for the license, excluding from any responsibility as to the use of the same by the candidate.

NLP TECHNICAL TEST

The goal of the test is working with a multi-language dataset, in order to demonstrate your Natural Language Processing and Machine Translation abilities. The Core Data Scientist and Storytelling attributes will also be evaluated during your resolution of the case.

About the Data:

The dataset you will be using is a multilingual, multi-context set of documents, which are a part of the one described on the following paper:

Ferrero, Jérémy & Agnès, Frédéric & Besacier, Laurent & Schwab, Didier. (2016). A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection.

Please note the dataset is divided on contexts/categories (Conference_papers, Wikipedia, …) and on languages, in the same way the folders are structured.

Objective 1: Create a document categorization classifier for the different contexts of the documents. You will be addressing this objective at context level, regardless of the language the documents are written in.

Tasks/Requirements:

- EDA: Exploratory data analysis of the Dataset.
- Reproducibility/Methodology: The analysis you provide must be reproductible. Your analysis will fulfill the Data Science methodology.
- Classification model: The deliverable will include a model which will receive a document as input and will output its class, which will be the context of that document.

Objective 2: Perform a topic model analysis on the provided documents. You will discover the hidden topics and describe them.

Tasks:

- Profile the different documents and topics.
- Provide a visualization of the profiles.

Documentation with the resolution of the case

Sending the file format to evaluate the result, preferable on **Github** platform.

Code used

The code used for the resolution of the case will be delivered in order to know the process used (methodology) and the degree of knowledge of the tool used for its resolution.

A **Storytelling** should also be provided explaining how the case (methodology) and conclusions of the result obtained have been made. The format of this point is free. The resolution of the case will then be explained in the personal technical interview.

Test run time: 1 week is sufficient.

Good Luck!!!