Technical assessment instruction

Language Detection

The aim of this assessment is to challenge you on a technical basis, involving machine learning, programming and research skills.

Goal

You goal is to develop an algorithm that detect the language of an arbitrary text (sentences or more lengthy texts). For example, your algorithm should return “en” for the English sentence “I am currently eating my breakfast”, or “fr” for the French sentence “J’ai oublié mon parapluie dans l’abribus”.

How to do it?

In order to carry out this assessment, we provide a dataset containing sentences in 21 languages (named “formatted_data”). This database is a .csv files whose separator is a “;”. The first column is the language code, the second one is the text, and the third one is the number of characters in the specified language.

This database is a collection of texts extracted from the European Union Proceedings.

A research article is provided to give you some insights on how to design such algorithm. We strongly advise you to use it in order to finish the assignment on time.

You are not required to use all languages to train your model and you can use any available resource to solve the problem (apart from directly using a pre-packaged language detector…).

No programming language is imposed for this assignment.

What do we expect from you?

- Correct implementation of the proposed model or another of your choice
- Good code quality (comments, overall clarity)
- An assessment of model performances and several ways to improve it
- Highlighting the limits of your model
- A live
    - demo of your code running on sentences examples (we can do it on Skype)
    - discussion about your work (Skype also)
- A very short “report” detailing your approach. In order to save time, we advise you to work on Jupyter Notebooks or any equivalent format in other languages. In that case, the live demo is embedded in the report.

Allowed time

This assignment should take several hours to complete (if you use the provided article). The maximum time you should spend on this problem is one full working day.

Deliverables

The expected deliverables are a "report" (as detailed in the "expectations" section) detailing your approach and your results, and your code.

They can be shared with us by email or through your GitHub account if you have one.

Assistance and feedback during the assignment

During this assessment you can contact me at:

- aurelien.baelde@upskills.ai
- aurelien.baelde@gmail.com during the weekend
- +33 6 23 20 62 73

if you need assistance (problem with data encoding for instance) or if instructions are not clear. You can also call me during the weekend (but I might not answer immediately).

Calling me is appreciated so don't hesitate if you want feedback or just bouncing ideas on this assignment.