07 - Conclusions & Future Work

August 2, 2020

1 EA Assignment 07 - Conclusions & Future Work

Authored by: Álvaro Bartolomé del Canto (alvarobartt @ GitHub)

1.1 Project Overview

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.

1.2 Project Conclusions

Both objectives have been successfully completed and their respective reports have been generated, tackling the problem as a Data Scientist should, including a detailed Story Telling on each research part developed. Additionally to the defined objectives, a detailed data exploration analysis and text preprocessing have been research/developed too, since it is probably the most relevant part of a NLP Data Scientist while tackling a NLP problem, as it is adding value to the raw data.

- `Objective 1`: the created model has been fit with 80% of the documents from every context and language and tested with the remaining 20% of the data with balanced contexts and languages too, achieving an accuracy of up to 98% on the validation set. Also this model has been dumped into a JOBLIB file so that it can be tested over unseen data.
- `Objective 2`: the topic modelling problem has been broken down into a topic modelling per context and language, so as to get more insights and analyse the hidden topics that can be found in each collection of documents, with also pretty satisfactory results evaluated in a supervised way.

To sum up, mention that even though the project tasks have been achieved and some extra points have been made, there is still much work ahead, so later on this Notebook, the Future Work will be defined.

1.3 Project Future Work

As Future Work, the main line of research should be focused on developing a consistent Machine Translation model in order to translate text from French and Spanish into English, which will indeed improve the results even though they are pretty accurate now.

Since in the first EA Interview with Francisco Martínez (EA Talent Coordinator) he spoke about the EA's project related to Machine Translation, it would make sense to proceed with the project designing a consistent Machine Translation model so as to test it's efficiency towards this problem.

Another Future Work line of research should be the design of Deep Learning models maybe in TensorFlow or PyTorch (usually more suitable for NLP), since we are presenting a simple use case along this project, but reality is a bit more complex, so tackling the problem using Deep Learning models should improve the model's performance when the input data is bigger, more contexts are provided and more languages too.

Finally, multilingual word embeddings should be used so as to improve the models performance whatever the input data is, so we should be using the word embeddings so as to "translate" every word in Spanish or French to English, so as to tackle the problem as a Multi-Lingual input one but for the model it would just be a single language. Also, when deploying the model into a production environment a reliable layer of language detection should be applied so as to either apply the word embeddings if the text is written in French or Spanish or discard the text if it is neither English, Spanish nor French.

1.4 Project Opinion

First of all, thank you for having come this far!

In my opinion the project suits perfectly to evaluate the NLP Data Scientist role (not just for EA but for every company) since a real use-case is presented with a strong research part, since the project is somehow based on a scientific publication, which motivates the candidate to research about it but keeping him/her motivated through the process, since as already mentioned, the problem could be a real use-case.

In my concrete case, before this project I never worked on multi-lingual problems even though I have worked on the whole NLP Pipeline applied to both English and Spanish, but not applied to French. Anyway, I found it really interesting since I had to do some research before tackling the Multi-Lingual Multi-Context Text Classification problem, but through the journey I have learnt new NLP concetps and adquired new NLP abilities, so either I get selected or not, at least I am proud to say that I have learnt something new.

When it comes to the results presented, I am pretty satisfied with the results I have obtained, mainly on the Text Classification model since I got to achieve a validation accuracy of over 98% which I think it is honestly pretty good with the data we have. By the other hand, I am not that satisfied with what I have done with the Topic Modelling part even though I got to achieve some pretty good results and a proper analysis of the identified profiles. Anyway, I want to keep on improving it, since even though I have spent a lot of hours trying to figure out how to tackle the problem so as to identify the best number of topics and profile them, I think that, with some more time, some parts can be updated.

Thank you again! I hope you take this into consideration as I would love to join the EA team as NLP Data Scientist.