00 - Project Definition

August 3, 2020

1 EA Assignment 00 - Project Definition

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 Analysis

So on, we need to create a text classification model so as to tackle the problem of classifying documents in their correct context, which is this case is sorted by the source they come from: Wikipedia, Conference Papers, APR (Amazon Product Reviews) and PAN11 (PAN-PC-11), where those documents are written in three different languages which are: English (en), French (fr) and Spanish (es).

This means that we will need to design a text classification model, which regardless of the language the documents are written in, is able to classify any document into the context it comes from.

So as to tackle this NLP problem, we will need to do a complete research detailing all the Data Science steps we need to complete during the problem solution so as to finally generate a model which does this classification.

Additionally, as the second objective we should also perform a topic modelling analysis using any topic modelling algorithm such as LDA (Latent Dirichlet Allocation), LSA/LSI (Latent Semantic Analysis/Indexing) or NMF (Non-Negative Matrix Factorization), which are un-supervised models used to automatically identify the different topics from a given collection of documents.

Note: every task should be properly documented in Jupyter Notebooks using a clear formatting and presenting every task in a proper Story Telling way.

1.3 Project Considerations

- Since the texts are written in multiple languages, either a multi-lingual preprocessing pipeline needs to be defined or, if that does not work, then a different preprocessing pipeline needs to be applied depending on the language which can be easily identified using any language detection Python library, so in this case we will need to define at least 3 different NLP PreProcessing Interfaces (one per language).
- The sources/contexts that the texts come from may contain additional stopwords context-based, which means that we will also need to clean the most frequent words not just the default stopwords from the listings. For example, in scientific publications some words such as Introduction, Abstract, Conclusions, Results, etc tend to be present in every scientific publication, so those words should be removed.
- As the dataset (available at https://www.dropbox.com/s/le9j5whzv3zzgrw/documents_challenge.zip?dl=0) contains a lot of texts, we will just need to properly define a preprocessing pipeline which preprocesses the text while it is being loaded so as to avoid multiple unnecessary FOR loops.
- To tackle the text classification model we will try out some scikit-learn model widely used for Text Classification such as Multinomial Naive Bayes or LinearSVC. Additionally, if the available resources support it, a Deep Learning framework as TensorFlow is suggested to be used since we have a multi-context multi-lingual dataset which means that the input shape of the data will be big and if the scikit-learn does not perform as well as expected, then the Deep Learning approach will be made.

- The main focus should be in the Story Telling part more than in the Text Classification one, since we want to extract useful conclusions so as to later improve the model, a perfect model lacking a proper Story Telling is hard to reproduce and scale.
- Every Jupyter Notebook should be reproducible, so absolute paths need to be avoided and all the managed data should be available in the GitHub repository.
- and more to come while some other considerations are made during the project's research!