Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220314

Streamlining Text Pre-Processing and Metrics Extraction

Elena Álvarez-García [a] and Daniel García-Costa [a] and Francisco Grimaldo [a,1]
[a] *Department of Computer Science, University of Valencia, Spain*

Abstract. Natural Language Processing involves reshaping and refining data sets into data that can be used for analysis, ensuring that the data is well formatted. The efficiency gap of data scientists spending most of their time preparing data is an opportunity for the technology sector to work on solutions to the problem. For this reason, a web tool has been developed that is capable of, on the one hand, speeding up the text cleaning process and, on the other hand, facilitating the extraction of metrics by analyzing and processing the texts through customized dictionaries in LIWC format, uploaded by the users themselves, and through sentiment analysis. All this, from a single interface that allows the user to customize the whole pipeline offering different modules for pre-processing and metrics extraction in order to be a solution to facilitate, streamline and automate the whole process.

Keywords. Text pre-processing, NLP, Text Mining, Metrics, Web tool

1. Introduction

Data quality is one of the main factors in data science, and clean data is important for creating great learning models and preparing data for latter analysis. The results of a CrowdFlower survey [2] of 16,000 data scientists reveal that time spent on pre-processing is one of the main obstacles. Compared to other tasks such as data mining and creating training sets, cleaning and organising data consumes around 60% of their time. According to Anaconda's 2020 annual study [1], the efficiency gap of data scientists spending most of their time preparing data for further analysis has been identified as an opportunity for the technology sector to work on solutions to the problem.

In Natural Language Processing (NLP), text processing refers to the practice of cleaning and preparing text data. To apply NLP techniques it is crucial that the corpus is well pre-processed. For example, data pre-processing is an essential step for the construction of machine learning models, whose outcomes can differ depending on it.

Pre-processing involves reshaping and refining texts into data that can be used for analysis, ensuring that the data is well formatted and follows a set of rules so that it can be understood and processed by machines [3]. Texts often contain characters such as punctuation or stop words that do not provide information and increase the complexity of the analysis. Thus, it is recommended to remove as much noise as possible before analyzing the data, in order to obtain clean data easy to process and to analyse later.

[1]Corresponding Author: Francisco Grimaldo, Department of Computer Science, University of Valencia, Spain; E-mail: francisco.grimaldo@uv.es.

Another relevant issue while analyzing text is text mining. It is based on different techniques and technologies that explore large amounts of data automatically or semi-automatically. Text mining seeks to discover information that was not present in a specific way, and text pre-processing is also a key step to successfully apply these techniques [6].

Nowadays there are multiple applications focused on text pre-processing (e.g. cleaning) and other focused on text mining (e.g. sentiment analysis), but they usually work independently. As for applications focused on text cleaning, two of the best known on the market are: Trifaca Wrangler [2] and OpenRefine [3] , both of them try to facilitate text pre-processing. The number of web tools capable of extracting sentiment from text is very high, being MonkeyLearn [4] one of the most renowned over the last year. This application hosts a set of text analysis tools, including sentiment analysis and keyword extraction. Unfortunately most of these applications are not free and do not provide a full set of tools that combine different types of techniques.

Unlike the aforementioned applications, the proposed web tool seeks to bring together both NLP techniques and text mining techniques using different Python libraries that allow the pre-processing of texts and the extraction of metrics for datasets in the simplest, fastest and most automatic way possible, having the aim of reducing the time spent in pre-processing and text mining, making data scientist's work easier and faster.

2. Proposal

The proposed web tool tries to make the pre-processing of text faster and to facilitate the extraction of metrics by applying dictionary based techniques, sentiment analysis and others, combining both methods for pre-processing and text mining in one single tool.

Bringing together these two types of techniques in a single tool will make it easier, for researchers not so familiar with specialised libraries, to perform these tasks in languages like Python, by offering an interface that allows them to automate and customise text pre-processing.

In order to obtain a scalable and modular web tool which can increase the number of features over time, it has been developed implementing different micro-services that provide each of the different functionalities.

The operation of the tool is straightforward. Users upload datasets in CSV format, they personalize the pipeline of the process and choose which order and characteristics to apply to each text column. As a result, the dataset is enriched with new columns depending on the customized pipeline. For example, if cleaning is the only technique selected, a new column with clean text is added, whereas if a custom dictionary extractor is activated, the user will get one new column for each category in the selected dictionary.

3. Use Case

By way of example, we present a use case of pipeline that provides the user with 4 different techniques, each one implemented in a different module (micro-service). Two of

[2] https://www.trifacta.com/
[3] https://openrefine.org/
[4] https://monkeylearn.com/

them are focused on text-preprocessing techniques (namely the Cleaning Module and the Translation Module) and the other two in text mining (namely the Custom Dictionary Module and the Emotions Module). The web tool then allows the user to configure the list of tasks and customize the characteristics of each module.

Cleaning Module: The aim of this module is to obtain clean texts that can be later parsed and analyzed without errors. Users can select which characteristics apply in this process such us: convert to upper or lower case, remove stop-words, remove punctuation marks, remove special characters, trim and remove multiple spaces or stem.

Translation Module: The purpose of this module is to translate texts into other languages. From the different Python libraries that perform this task, we use Google Trans, the free Python library that implements the Google API. This library works by making calls to the Google Translate API and allows to translate texts into all the languages supported by the Google API.

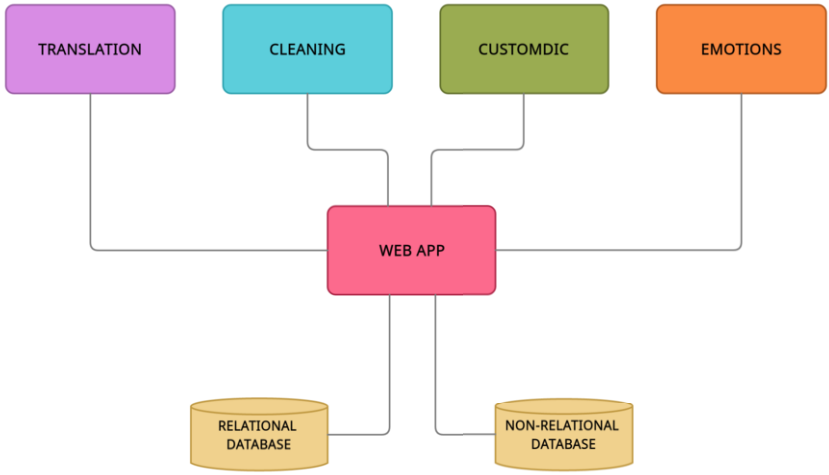
Custom Dictionary Module: This module is responsible for making quantitative analysis of texts by checking for the presence of words given by a dictionary in the LIWC [5] (Linguistic Inquiry and Word Count) format uploaded by the user. LIWC is a well-known format available for a large number of existing dictionaries and its ease to use by people with less specialised knowledge. Using LIWC, words can be counted and grouped in meaningful categories [5]. In this way, it is possible to obtain a value for each of the dimensions defined by the dictionary. These metrics are extracted by means of the liwc[6] library in Python, which parses dictionaries and counts the occurrences of words within the text. We then obtain the percentage with respect to the total number of words in the text. The outcome of this module adds a new column to the dataset for each of the categories specified in the dictionary.

Emotions Module: Sentiment analysis, also known as opinion mining, is the field of study that analyses people's opinions, feelings, evaluations, attitudes and emotions towards entities and their attributes expressed in written text [4]. This module extracts sentiments form texts written in English and Spanish using the libraries Lexmo [7] and pysentimiento [8], respectively.

The following figure shows the architecture of the web tool using the 4 modules previously explained. Each of these modules is an independent micro-service that communicates with the application that displays the web interface and from which the user interacts. This web application is in charge of the persistence of all the data, using a relational database for user information and a non-relational one for storing the datasets. The use of a non-relational database allows us to deal with different data structures in the provided datasets.

[5] https://www.liwc.app/
[6] https://github.com/chbrown/liwc-python
[7] https://github.com/dinbav/LeXmo
[8] https://github.com/pysentimiento/pysentimiento

Figure 1. Application architecture



4. Conclusions

The tool allows the application of NLP and text mining techniques such as cleaning and extraction of metrics such as sentiment analysis through an easy-to-use interface, obtaining machine and human understandable information. The entire processing pipeline can be customised by the user, helping scientists from other disciplines that do not have much knowledge in these these pre-pocessing, and also for more experienced people to speed up certain tests or parts of their work.

The application is currently in alpha phase and is expected to be made accessible to anyone when it is in a more stable state.

References

[1] Anaconda. Moving from hype toward maturity 2020 state of data science, 2020.

[2] CrowdFlower. Datascience report, 2016.

[3] Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, Freeson Kaniwa, and Kutlwano Sedimo. Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, 6(3):207–210, 2016.

[4] Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38, 2015.

[5] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[6] Yu Zhang, Mengdong Chen, and Lianzhong Liu. A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 681–685, 2015.