An approach for the analysis of news during COVID-19 in the Chubut province

Pablo Toledo Margalef[1], Emanuel Balcazar[3], Leo Ordinez[2], Claudio Delrieux[2], and Lucila Allende[3]

[1] Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH)- CCT-CENPAT -CONICET, Puerto Madryn, Argentina
ptoledo@cenpat-conicet.gob.ar
[2] Laboratorio de Investigación en Informática (LINVI), Facultad de Ingeniería, UNPSJB, Puerto Madryn, Argentina
[3] Facultad de Ingeniería, UNPSJB

Abstract. The present work exposes preliminary results on utilizing web scraping and data mining techniques to analyze news articles published during the COVID-19 pandemic in the Chubut province. Analysis of extracted articles was made using Latent Dirichlet Allocation obtaining promising results.

Keywords: LDA · COVID-19 · media · news · NLP · scraping

1 Introduction

This work is framed in a project which aims at constructing knowledge in order to evaluate the current and predict the future socio-economic situation of the Chubut province. In particular, focusing on vulnerable population in the context of pandemic generated by the COVID-19. The target region is limited by the geographical limits of the Chubut province, adjusting the territorial scale to cities, towns and rural communes. The knowledge part of interest for this article will be constructed by extracting, processing, and automatically analyzing news articles published in local press with provincial scope. The goal is to evaluate the evolution of different topics that affects the community. Results will be obtained through web scraping and the application of Natural Language Processing Techniques (NLP).

The main objective of the current work is the construction of knowledge about the current situation of the Chubut province regarding the COVID-19 pandemic through extraction and analysis of news around topics affected by the sanitary context. The use of an unsupervised technique, like Latent Dirichlet Allocation (LDA), leads us to a discovery of such topics instead of a confirmation about preestablished subjects.

2 Materials and Methods

For the analysis of news LDA is used. This technique is part of Natural Language Processing (NLP) [8], and is considered an Unsupervised Learning method [4].

In the case of NLP, if the observations are words collected into documents, the model posits that each document is a mixture of topics which can be attributable to the presence of certain terms and that each word's presence can be attributable to one or more of the document's topics [2].

The media chosen for extraction were only those from the Chubut province. This was to limit the number of results and focus particularly in that region. The selection was also influenced by the popularity of their web portal, the amount of news posted and the territorial representativeness such that the complete provincial territory is included.

The extraction process was divided into steps to facilitate development of different components, each with its own responsibility within the extraction and analysis of articles. The steps are:

1. **Google Search:** as a first step, the search equations for each site are retrieved from the database. An equation indicates the date and URL to be used for the search. Then a personalized Google search engine performs the actual search and obtains the URLs for the links meeting the criteria.
2. **Extraction:** the links obtained in the previous step are received by a component whose task is to extract the articles in HTML format.
3. **Cleaning:** The HTML is processed extracting the relevant sections, such as the title, subtitle, body, date and original link. This way we obtain a clean version of the article.
4. **Normalization:** The articles are normalized in a common format to store them in the database, this eases the creation of a unified dataset across all sites used.
5. **NLP:** Natural Language Processing is applied once the articles are stored in the database. A series of modifications are made to the articles so they are useful in future processing, each term is taken to its root and unnecessary words (*i.e.*, stop words) are removed.

In Fig. 1 the scraping process is sketched. Note that step 3. implicitly establishes a *plugin* architecture since each website has a different HTML structure.

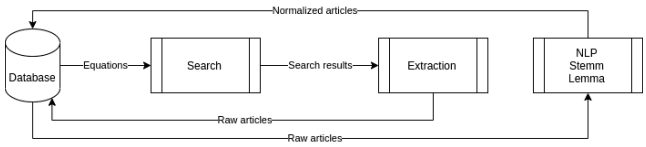

Fig. 1: Web scraping process of news media.

The tools used in the previous steps were developed in NodeJS, using AdorniJS for one of the components. Python as NLP module, and PostgreSQL to store extraction a post processing results. Rabbit MQ was used for communication between each component.

Extracted articles have the following structure:

- `title`: title of the article.
- `snippet`: brief article summary obtained through Google.
- `link`: original article link.
- `displayLink`: base URL of the site where the article was extracted from.
- `body`: body of the article, most important field in the dataset.
- `published`: date the article was published on.
- `expected_date`: expected publication date. It is used to control whether the article corresponds to the date of the search equation used.
- `is_useful`: Indicates whether the article is useful for processing. It is not field in use.
- `analyzed`: Indicates if the article is already processed by NLP.

From March 2020 to March 2021, more than 62,000 items were obtained, from which almost 85,000 unique words were found. Fig. 2 shows the dashboard of the application developed for the extraction of news. The system keeps on extracting data, so this numbers are expected to increase over time. The finishing date of this process is yet to be decided. In the figure, the top 5 of recurrent words and the different news sites extracted along with their amount of articles, are depicted. The most recurring words are: *province, case, work, do /make* (in Spanish, *hacer*) and *power* (it can also be, "can", since the word was *poder).*

Fig. 2: Screenshot of the application developed for the scraping.

To process data and generate the corresponding models, an implementation in Python was performed. The pre-processing of the documents was made using *Spacy*[4], an NLP library providing functionalities to generate the root of the terms, being it lemmatization or stemming. To generate the models we used *Gensim*, this library allows the construction of topic models through various methods, including LDA[9]. Visualizations were possible through *pyLDAvis*[10]. Lastly the retrieving of the data from storage mediums and the raw data processing was possible using data science tools such as *pandas*[7] and *numpy*[3].

[4] `https://nightly.spacy.io/`

Applying these methods to news is an insightful technique for knowledge construction. Previous applications can be found on finance articles [5], detecting risk of a pandemic [1], or even analyzing patters within media coverage of health communications on early stages of the COVID-19 outbreak[6].

3 Results

Using a data set of 1,103 articles, a series of models were constructed. At first lemmatization and stemming was not used, and 15, 30 and 60 topics were extracted. It was observed that the resulting topics were widely spaced from each other or overlapped. There was no uniformity between the relevance of each topic, resulting in topics excessively larger than others. After lemmatization and stemming were applied, keeping the quantities of extracted topics, no improvement was observed. Figures 3a 3b respectively show the previous situations for the case of 30 topics.

One last experiment was conducted using only articles from a single day, but this time only 6 topics were extracted, applying lemmatization and stemming. A clear separation between topics and uniformity among the topic sizes was observed (see Fig. 3c).

A visualization of the described models can be found at the following link[5].

(a) 30 topics with no preprocessing

(b) 30 topics with lemmatization

(c) 6 topics from a single day with lemmatization

Fig. 3: LDA experiments.

4 Conclusions and Future Work

This ongoing research showed the potential in analyzing news articles for the understanding of a complex phenomenon such as that experienced by the COVID-

[5] https://papablo.gitlab.io/resultados-short-paper-analisis-noticias-chubut/

19 pandemic. The automatic extraction of information and its analysis using techniques such as NLP showed their potential in terms of quantitative studies.

It was possible to discern a reasonable number of topics to be extracted that allowed a similar size and separation between them. However, these results were obtained using articles from a single day. Future works are to be focused on considering time as another analysis dimension in order to study the dynamic evolution of topics and terms. This would allow us to generate a hypothesis and list of terms to be followed over time.

References

1. Akrouchi, M.E., Benbrahim, H., Kassou, I.: End-to-end LDA-based automatic weak signal detection in web news. Knowledge-Based Systems **212**, 106650 (Jan 2021). https://doi.org/10.1016/j.knosys.2020.106650
2. Blei, D., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
3. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
4. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (2019)
5. Kakhki, S.S.A., Kavaklioglu, C., Bener, A.: Topic detection and document similarity on financial news. In: Advances in Artificial Intelligence, pp. 322–328. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-89656-4_34
6. Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C.J.P., Ming, W.K.: Health communication through news media during the early stage of the COVID-19 outbreak in china: Digital topic modeling approach. Journal of Medical Internet Research **22**(4), e19118 (Apr 2020). https://doi.org/10.2196/19118
7. McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference. pp. 51 – 56 (2010)
8. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association **18**(5), 544–551 (09 2011). https://doi.org/10.1136/amiajnl-2011-000464
9. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), `http://is.muni.cz/publication/884893/en`
10. Sievert, C., Shirley, K.E.: Ldavis: A method for visualizing and interpreting topics. pp. 63–70. Baltimor, Maryland, USA (June 2014)