Webscraping plan

Jorge Roa

Context

During the first year or two of the pandemic, political heads and head experts gave (in some cases daily) press conferences, explaining the situation and their strategies. In the US for example, Anthony Fauci (director of NIH's National Institute of Allergy and Infectious Disease) was helping US government to implement different strategies to fight COVID-19 pandemic. In Mexico, Hugo López-Gatell (Undersecretariat of Prevention and Health Promotion at the Mexican Secretariat of Health) played a major role as a spokesman and one of the lead members of the task force addressing the COVID-19 pandemic. In general, every country posses a team of experts that deal with the pandemic, interacting with the public and the media.

Objective

Create an "expert" corpus from the Spanish authorities that were on duty during the COVID-19 pandemic. For this, we will first identify the key stakeholders that were in charge in taking decisions and speaking in public related to the pandemic. Once, we define these actors, we will retrieve data to construct an "expert" corpus containing the links, the speaker, the date of the declaration and the corpus speech of this experts. This data will be retrieve starting from January 2020 till March 2023.

How we will retrieve these corpus, speeches and texts?

We will implement different approaches, but the main will be through webscraping using `BeautifulSoup`. Also, we will implement some repositories that contain corpus of spanish political speeches from "El Congreso" that are updated till the final months of 2020.

Key Stakeholders

Pedro Sánchez
President of Spain

Margarita Robles
Minister of Defense of Spain

Salvador Illa
Minister of Health (2020-2021)

Carolina Darias
Minister of Health (2021-2023)

Fernando Simón
Director of the Coordination Center for Health Alerts and Emergencies

José Luis Ábalos
Head of Transport, Mobility and Urban Agenda

Salvador Illa
Minister of Health (2020-2021)

Carolina Darias
Minister of Health (2021-2023)

Pero Godoy
President of the Spanish Society of Epidemiology

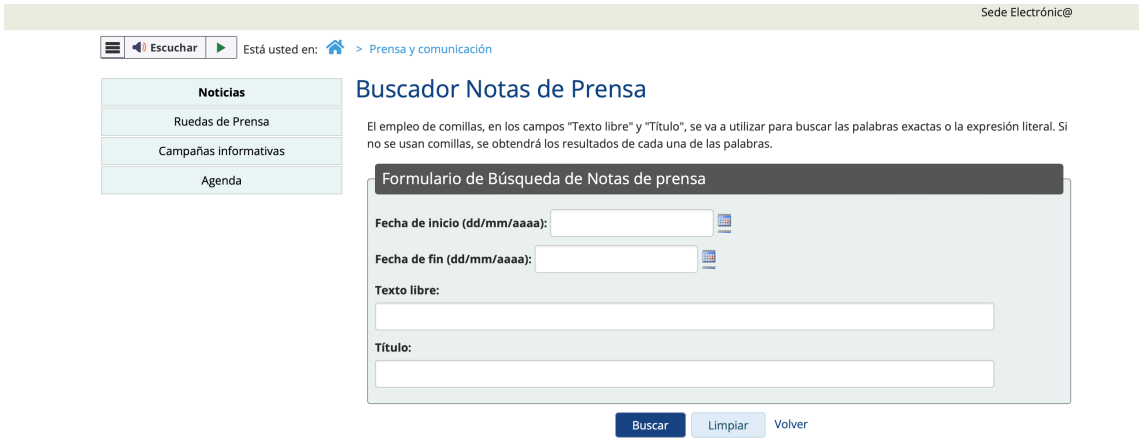
These are just one of the most important stakeholders during the COVID-19 pandemic in Spain. However, during the retrieve of the speeches and texts, we could add more actors to this list, depending the availability of information.

Data

Ministerio de Sanidad.- Notas

Ministerio de Sanidad

Press notes from official sources such as the Ministery of Health are perfect to portrait the position of the different stakeholders involved in the pandemic. In this case, we are expecting to have more notes related to Salvador Ila and Carolina Darias.



The idea is use keywords and filter them with dates between the pandemic started (January 2020) till today. Some of this press conferences are about news, but also about declarations of different stakeholders. Some of them are in a pdf format. For this, we are planning to retrieve that pdf's to extract the text of them and put it in a csv file. If that is not the case, we will handle the text and put it directly into the csv file.

Press releases related to the new coronavirus, COVID-19

Press releases related to COVID-19

This is tool is more specific and it comes from the same source of the Ministery of Health. Here the research will be directly since we will retrieve every press conference and evaluate if it contains statements or discourses from stakeholders.

La Moncloa: Official website of the Government

This website will be one of our main sources to retrieve the speeches. The website allow us to divide the statements and speeches between positions. For example, we can search statements and speeches that are only from the president. Also, we can do the same with the Council of Ministers like Salvador Illa or Carolina Darias.

President

Council of Ministers

For different research, this source has been used for other scholars to do their research. For example:

Tip

Montiel, C. J., Uyheng, J., & Dela Paz, E. (2021). The Language of Pandemic Leaderships: Mapping Political Rhetoric During the COVID-19 Outbreak. Political psychology, 42(5), 747–766. https://doi.org/10.1111/pops.12753

Congreso de los diputados

The Chamber of Deputies website registers all the interventions of the deputies. In this context, we will use keywords to retrieve interventions of our stakeholders of interest. to retrieve this data, `BeautifulSoup` will be used too. Fortunately, there is one repo that contains different interventions in the Chamber, so, that database will be inspected and we will retrieve our targets.

Congreso de los diputados

GitHub repository: 2020

Parliamentary corpora

Parliamentary corpora are a very important multidisciplinary language resource that can be approached from many research perspectives, including not only political science, but also sociology, history, psychology, and applicative approaches to linguistics, for instance, critical discourse analysis. The good availability of parliamentary proceedings in digitized form and granted access rights to public information in the EU countries have motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora.

This project is super interesting since it holds different types of databases to use this text as inputs for NLP models. For our case, we can use a database that contains

Go to the database and the information related to it.

Here is more information about whta the database ParlaMintES contains

It can be helpful too since it's already structured in the format that we are looking for with extra information like type of parlament and type of intervention.

For comments, changes of the existing content or inclusion of new corpora, send us an email.

This website was last updated on 24 October 2022.

Parliamentary corpora in the CLARIN infrastructure

Corpus	Language	Description	Availability
Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1 **Size:** 3.7 million utterances, 495 million words **Annotation:** tokenised, MSD-tagged, syntactically parsed (Universal Dependencies) **Licence:** CC BY 4.0	Bulgarian, Croatian, Czech, Danish, Dutch, English, French, Hungarian, Icelandic, Italian, Latvian, Lithuanian, Polish, Slovenian, Spanish, Turkish	ParlaMint is a multilingual set of comparable corpora containing parliamentary debates mostly starting at the end of 2015 and extending to mid 2020, with each corpus being about 20 million words in size. The sessions in the corpora are marked as belonging to the COVID-19 period (after October 2019), or being "reference" (before that date). The corpora have extensive meta-data about the speakers (name, gender, party affiliation, MP status), are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches also contain marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc. The corpus is available for download from the CLARIN.SI repository and through the concordancer noSketchEngine. Note that the version of the corpus without linguistic mark-up is available for download under a separate CLARIN.SI entry.	Concordancer Download

Regarding the use of open sources as press and media, we have the next option.

BONUS: Using the GoogleNews package

Finally, we have this alternative option for retrieve text. This code was used to do research about criminal groups in Mexico and their presence reported in news and press. We need to be aware that this is not the most precise tools since also pulls news that are not related for the topic of research. However, it's worth a shot since it can give us the date, the source and the title where it retrieved the information and text.

```
pip install GoogleNews
pip install newspaper3k
pip install pandas
pip install nltk
pip install openpyxl
```

```
from GoogleNews import GoogleNews
from newspaper import Article
from newspaper import Config
import pandas as pd
import nltk

nltk.download('punkt')

user_agent = 'your_user_agent'
config = Config()
config.browser_user_agent = user_agent

googlenews=GoogleNews(start='01/03/2020',end='02/14/2021')
googlenews.search('Sedena presuntos')
result=googlenews.result()
df=pd.DataFrame(result)
print(df.head())

print(df)

for i in range(2,5):
    googlenews.getpage(i)
    result=googlenews.result()
    df=pd.DataFrame(result)
list=[]
for ind in df.index:
    dict={}
    article = Article(df['link'][ind],config=config)
    article.download()
    article.parse()
    article.nlp()
    dict['Date']=df['date'][ind]
    dict['Media']=df['media'][ind]
    dict['Title']=article.title
    dict['Article']=article.text
    dict['Summary']=article.summary
    dict['url']=article.url
    list.append(dict)
news_df=pd.DataFrame(list)
news_df.to_excel("prueba.xlsx")
```