Taller 2 - Web Scrapping

Daniel Alejandro Chimbi León

Edna Valentina Henao Barrera

Daniel Felipe Sanchez Mogollón

Facultad de Ingeniería, Universidad El Bosque

Big Data Analytics

Fabian Camilo Peña Lozano

22 de Octubre de 2021

INTRODUCCIÓN

El Web Scrapping es una técnica para poder extraer y almacenar los datos de las páginas web con el fin de analizar y utilizar su contenido, a través del web scrapping se pueden almacenar varios tipos de información, para nuestra práctica sería la información de los artículos y posteriormente alojar dicha información en Mongo DB.

En el presente taller se realizará web scrapping de forma automática, ya que se recurre a los algoritmos para poder analizar las diferentes páginas web para extraer la información, dentro del análisis automático existen 3 tipos de modos para realizarlos, los cuales son: Analizador sintáctico, bots y texto.

El modo que se utilizará para el taller es el de analizador sintáctico o también llamados *"parsers"*, ya que se utilizan para poder convertir un texto en una estructura nueva, en este caso HTML y posteriormente a ello, almacenar la información.

PARTE 1

Para empezar es esencial instalar selenium y pymongo, para poder realizar el proceso de web scrapping y almacenar los datos obtenidos de cada artículo.

Instalación de selenium.

```
PS C:\Users\ASUS> pip install selenium
Collecting selenium
  Downloading selenium-4.0.0-py3-none-any.whl (954 kB)
     |████████████████████████████████| 954 kB 3.3 MB/s
Collecting trio~=0.17
  Downloading trio-0.19.0-py3-none-any.whl (356 kB)
     |████████████████████████████████| 356 kB 6.4 MB/s
Collecting trio-websocket~=0.9
  Downloading trio_websocket-0.9.2-py3-none-any.whl (16 kB)
Requirement already satisfied: urllib3[secure]~=1.26 in c:\users\asus\anaconda3\lib\site-packages (from selenium) (1.26.4)
Requirement already satisfied: async-generator>=1.9 in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.10)
Requirement already satisfied: idna in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.10)
Requirement already satisfied: sniffio in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.2.0)
Collecting outcome
  Downloading outcome-1.1.0-py2.py3-none-any.whl (9.7 kB)
Requirement already satisfied: attrs>=19.2.0 in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (20.3.0)
Requirement already satisfied: cffi>=1.14 in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.14.5)
Requirement already satisfied: sortedcontainers in c:\users\asus\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.3.0)
Requirement already satisfied: pycparser in c:\users\asus\anaconda3\lib\site-packages (from cffi>=1.14->trio~=0.17->selenium) (2.20)
Collecting wsproto>=0.14
  Downloading wsproto-1.0.0-py3-none-any.whl (24 kB)
Requirement already satisfied: cryptography>=1.3.4 in c:\users\asus\anaconda3\lib\site-packages (from urllib3[secure]~=1.26->selenium) (3.4.7)
Requirement already satisfied: pyOpenSSL>=0.14 in c:\users\asus\anaconda3\lib\site-packages (from urllib3[secure]~=1.26->selenium) (20.0.1)
Requirement already satisfied: certifi in c:\users\asus\anaconda3\lib\site-packages (from urllib3[secure]~=1.26->selenium) (2020.12.5)
Requirement already satisfied: six>=1.5.2 in c:\users\asus\anaconda3\lib\site-packages (from pyOpenSSL>=0.14->urllib3[secure]~=1.26->selenium) (1.15.0)
Collecting h11<1,>=0.9.0
  Downloading h11-0.12.0-py3-none-any.whl (54 kB)
     |████████████████████████████████| 54 kB 1.5 MB/s
Installing collected packages: outcome, h11, wsproto, trio, trio-websocket, selenium
Successfully installed h11-0.12.0 outcome-1.1.0 selenium-4.0.0 trio-0.19.0 trio-websocket-0.9.2 wsproto-1.0.0
```

Instalación de pymongo

```
PS C:\Users\ASUS> pip install pymongo
Collecting pymongo
  Downloading pymongo-3.12.1-cp38-cp38-win_amd64.whl (398 kB)
     |████████████████████████████████| 398 kB 3.3 MB/s
Installing collected packages: pymongo
Successfully installed pymongo-3.12.1
```

Luego de instalar las respectivas librerías, se realizan los siguientes pasos:

1. Se importan las librerías previamente instaladas.
2. Se crea la conexión con MongoDB para almacenar la información extraída del web scrapping.
3. Se especifica la url de la página que se analizará.
4. Se establece el driver según el sistema operativo, en este caso, el ejecutable para Windows "geckodrive.exe", el cual se debe ubicar en el espacio de trabajo en el entorno Jupyter.
5. Se utiliza el driver para poder abrir una nueva ventana del navegador Firefox.

[1]:
```python
# Importing required libraries

import time
import datetime

from bs4 import BeautifulSoup
from selenium import webdriver

from pymongo import MongoClient
```

[2]:
```python
# Creating a connection to MongoDB
client = MongoClient('localhost', 27017)
db = client['news']
collection = db['elespectador']
```

[3]:
```python
# Base URL of the site to be analyzed
SITE_URL = 'https://www.elespectador.com'
```

[4]:
```python
# Firefox web driver path
# Download the driver for you S.O. here: https://github.com/mozilla/geckodriver/releases
DRIVER_PATH = './geckodriver.exe'
```

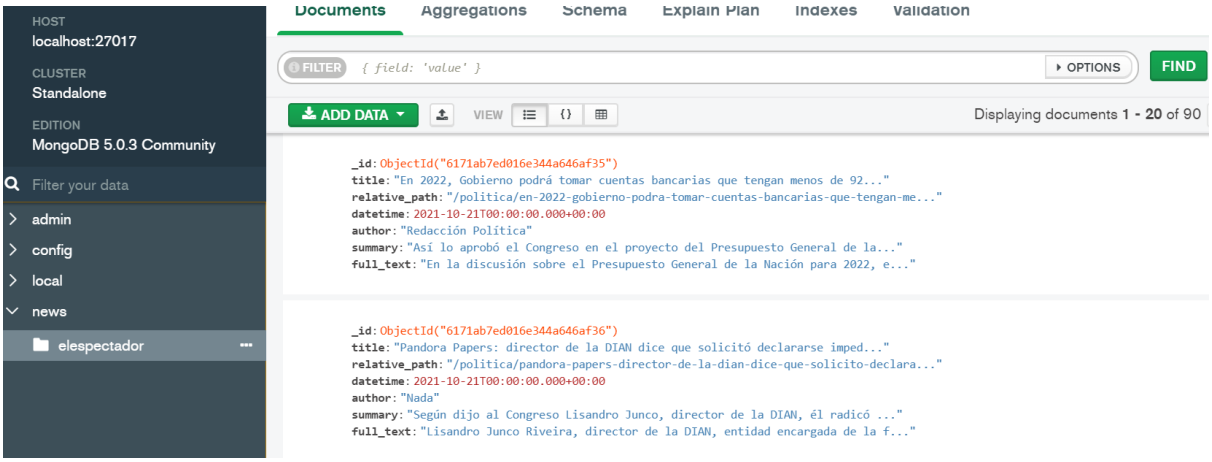
[5]:
```python
# Creating a new firefox window
browser = webdriver.Firefox(executable_path = DRIVER_PATH)
```

Para poder analizar la página en las 5 categorías distintas, se debe realizar de manera dinámica, por ende, se realiza un método que recorra la dichas categorías en al menos 3 páginas de cada una.

```
mayor = ['archivo']
categoria = ['politica', 'ciencia', 'salud', 'educacion', 'judicial']
paginacion = [1, 2, 3]


for m in mayor:
    for c in categoria:
        for p in paginacion:
            soup = make_request(browser, '/'+m+'/'+c+'/'+str(p)+'/')
            # Finding the section where news are contained
            layout = soup.find(class_ = 'Layout-flexAds')
            # Getting blocks from layout
            blocks = layout.find('section').find_all(recursive = False)
            print(len(blocks)) # 3 blocks founded, 2 for news and 1 for pagintion
            # Finding and concatenating news cards
            cards = blocks[0].find_all(class_ = 'Card') + blocks[1].find_all(class_ = 'Card')
            len(cards)
            news = []

            for card in cards:
                news.append({
                    'title': card.find('h2', class_ = 'Card-Title').find('a').get_text(),
                    'relative_path': card.find('h2', class_ = 'Card-Title').find('a')['href']
                })

            news
            for n in news:
                # Getting HTML content for news page
                soup = make_request(browser, n['relative_path'])

                # Extracting news metadata

                if soup.find(class_ = 'ArticleHeader-Date'):

                    var = soup.find(class_ = 'ArticleHeader-Date').get_text()
                    x = []
                    x = var.split('-')
                    newvar = ''.join(char for char in x[0] if char.isalnum())
                    var2 = newvar.replace("sept","sep")
                    n['datetime'] = datetime.datetime.strptime(var2, "%d%b%Y")
                    #n['datetime'] = soup.find(class_ = 'ArticleHeader-Date').get_text()
                else:
                    n['datetime'] = 'Nada'
```

```python
    if soup.find(class_ = 'ACredit-Author').find('a'):
        n['author'] = soup.find(class_ = 'ACredit-Author').find('a').get_text()
    else:
        n['author'] = 'Nada'

    if soup.find(class_ = 'ArticleHeader-Hook').find('div'):
        n['summary'] = soup.find(class_ = 'ArticleHeader-Hook').find('div').get_text()
    else:
        n['summary'] = 'Nada'


    # Extracting and concatenating news full text
    paragraphs = soup.find_all(class_ = 'font--secondary')
    n['full_text'] = ' '.join([p.get_text() for p in paragraphs])
news
# Storing extracted information for further analysis
collection.insert_many(news)
```

Al finalizar podemos evidenciar que la información de las noticias se almacena correctamente con sus respectivos atributos en MongoDB.



PARTE 2

Realizamos la respectiva instalación de wordcloud para poder crear nube de palabras con el dataset generado en la parte 1

```
[1]: !pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\users\asus\anaconda3\lib\site-packages (1.8.1)
Requirement already satisfied: numpy>=1.6.1 in c:\users\asus\anaconda3\lib\site-packages (from wordcloud) (1.20.1)
Requirement already satisfied: matplotlib in c:\users\asus\anaconda3\lib\site-packages (from wordcloud) (3.3.4)
Requirement already satisfied: pillow in c:\users\asus\anaconda3\lib\site-packages (from wordcloud) (8.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\asus\anaconda3\lib\site-packages (from matplotlib->wordcloud)
(2.4.7)
Requirement already satisfied: cycler>=0.10 in c:\users\asus\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\asus\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.3.1)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\asus\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: six in c:\users\asus\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib->wordcloud) (1.15.0)
```

Realizamos las diferentes importaciones de las librerías necesarias para el análisis de texto.

```
import os
import time
import datetime
import nltk

from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords
from nltk.tag import StanfordPOSTagger


from wordcloud import WordCloud

from pymongo import MongoClient
from pymongo import TEXT

import matplotlib.pyplot as plt
%matplotlib inline

nltk.download('punkt')
nltk.download('stopwords')

import re
```

Nuevamente nos conectamos a la base de datos, para poder acceder a los datos de nuestro dataset.

```
# Creating a connection to MongoDB
client = MongoClient("localhost", 27017)
db = client["news"]
collection = db["elespectador"]
```

Recorremos los atributos de cada uno de nuestros objetos de nuestro dataset y seleccionamos el título, el resumen y el texto completo para posteriormente imprimirlo y evidenciar que se une en un texto grande para poderlo analizar.

```
[4]: text = []
     for news in list(collection.find({}, {"title": 1, "summary": 1, "full_text": 1, "_id": 0})):
         text.append(news["title"])
         text.append(news["summary"])
         text.append(news["full_text"])
```

```
[5]: text = " ".join(text)
```

```
[6]: text
```

El resultado es el siguiente.

```
[6]: 'También para Congreso: ya van 129 comités que buscan recolectar firmas La Registraduría precisó que no se exigirán más de 50.000 firmas válidas
para la inscripción con miras a las elecciones de marzo próximo. El plazo para el registro de comités vence el 13 de noviembre del 2021. No solo
es un fenómeno que se presente con miras a las elecciones para presidente de la República. Según cifras de la Registraduría Nacional reveladas es
te jueves, hasta ahora se han registrado 129 comités inscriptores al Congreso por grupos significativos de ciudadanos y movimientos sociales, bus
cando avalar sus candidaturas a través de firmas.  En contexto: ¿Campaña anticipada? Póngale la firma De acuerdo con la entidad, los comités insc
ritos no son solo para avalar candidaturas, sino para promover el voto en blanco con miras a las elecciones legislativas, que se celebrarán en ma
rzo próximo. “Del total, 66 corresponden a comités inscriptores de candidaturas a la Cámara de Representantes (61 por circunscripción territorial
y cinco por la circunscripción internacional), 41 a comités inscriptores de candidaturas al Senado, 8 a comités promotores del voto en blanco (cu
atro por Cámara y cuatro por Senado) y 14 a comités que desistieron del registro (nueve a la Cámara y cinco al Senado”, explicó la Registraduría.
En el caso de los comités para Congreso, deberán recoger un número mínimo de firmas válidas equivalente al 20 % del resultado de la división de l
a cantidad de ciudadanos aptos para votar en la respectiva circunscripción electoral, entre el número de curules por proveer. En todo caso, en ni
ngún caso se exigirán más de 50.000 firmas válidas para la inscripción. Lea también: Firmas: ¿espejismo o avance democrático? “El término para el
registro de los comités inscriptores de candidaturas y promotores del voto en blanco apoyados por grupos significativos de ciudadanos y movimient
os sociales vence el 13 de noviembre del 2021, esto, de acuerdo con lo establecido en el calendario electoral para las elecciones del Congreso”,
precisó. Con corte al pasado 17 de septiembre, según datos de la Registraduría conocidos por este diario, solo para Presidencia de la República s
e contaba con al menos 30 comités para recoger las firmas que servirán de sostén para el aspirante. Ello implica que, en solo cuatro meses, y a t
res de que concluya la inscripción de más  ya se cuentan más de la mitad del total de grupos que se registraron para 2018  Los beneficios de conf
```

El texto se cambia a minúscula.

```
[7]: text = text.lower() # WARNING
```

Se realiza la tokenización y se imprime el resultado.

```
[9]: words = word_tokenize(text)

     print(words)
```

```
['también', 'para', 'congreso', ':', 'ya', 'van', '129', 'comités', 'que', 'buscan', 'recolectar', 'firmas', 'la', 'registradur
ía', 'precisó', 'que', 'no', 'se', 'exigirán', 'más', 'de', '50.000', 'firmas', 'válidas', 'para', 'la', 'inscripción', 'con',
'miras', 'a', 'las', 'elecciones', 'de', 'marzo', 'próximo', '.', 'el', 'plazo', 'para', 'el', 'registro', 'de', 'comités', 've
nce', 'el', '13', 'de', 'noviembre', 'del', '2021.', 'no', 'solo', 'es', 'un', 'fenómeno', 'que', 'se', 'presente', 'con', 'mir
as', 'a', 'las', 'elecciones', 'para', 'presidente', 'de', 'la', 'república', '.', 'según', 'cifras', 'de', 'la', 'registradurí
a', 'nacional', 'reveladas', 'este', 'jueves', ',', 'hasta', 'ahora', 'se', 'han', 'registrado', '129', 'comités', 'inscriptore
s', 'al', 'congreso', 'por', 'grupos', 'significativos', 'de', 'ciudadanos', 'y', 'movimientos', 'sociales', ',', 'buscando',
'avalar', 'sus', 'candidaturas', 'a', 'través', 'de', 'firmas', '.', 'en', 'contexto', ':', '¿campaña', 'anticipada', '?', 'pón
gale', 'la', 'firma', 'de', 'acuerdo', 'con', 'la', 'entidad', ',', 'los', 'comités', 'inscritos', 'no', 'son', 'solo', 'para',
'avalar', 'candidaturas', ',', 'sino', 'para', 'promover', 'el', 'voto', 'en', 'blanco', 'con', 'miras', 'a', 'las', 'eleccione
s', 'legislativas', ',', 'que', 'se', 'celebrarán', 'en', 'marzo', 'próximo', '.', '“', 'del', 'total', ',', '66', 'corresponde
n', 'a', 'comités', 'inscriptores', 'de', 'candidaturas', 'a', 'la', 'cámara', 'de', 'representantes', '(', '61', 'por', 'circu
nscripción', 'territorial', 'y', 'cinco', 'por', 'la', 'circunscripción', 'internacional', ')', ',', '41', 'a', 'comités', 'ins
criptores', 'de', 'candidaturas', 'al', 'senado', ',', '8', 'a', 'comités', 'promotores', 'del', 'voto', 'en', 'blanco', '(',
'cuatro', 'por', 'cámara', 'y', 'cuatro', 'por', 'senado', ')', 'y', '14', 'a', 'comités', 'que', 'desistieron', 'del', 'regist
ro', '(', 'nueve', 'a', 'la', 'cámara', 'y', 'cinco', 'al', 'senado', '”', ',', 'explicó', 'la', 'registraduría', '.', 'en', 'e
l', 'caso', 'de', 'los', 'comités', 'para', 'congreso', ',', 'deberán', 'recoger', 'un', 'número', 'mínimo', 'de', 'firmas', 'v
álidas', 'equivalente', 'al', '20', '%', 'del', 'resultado', 'de', 'la', 'división', 'de', 'la', 'cantidad', 'de', 'ciudadano
s', 'aptos', 'para', 'votar', 'en', 'la', 'respectiva', 'circunscripción', 'electoral', ',', 'entre', 'el', 'número', 'de', 'cu
rules', 'por', 'proveer', '.', 'en', 'todo', 'caso', ',', 'en', 'ningún', 'caso', 'se', 'exigirán', 'más', 'de', '50.000', 'fir
mas', 'válidas', 'para', 'la', 'inscripción', '.', 'lea', 'también', ':', 'firmas', ':', '¿espejismo', 'o', 'avance', 'democrát
ico', '?', '“', 'el', 'término', 'para', 'el', 'registro', 'de', 'los', 'comités', 'inscriptores', 'de', 'candidaturas', 'y',
'promotores', 'del', 'voto', 'en', 'blanco', 'apoyados', 'por', 'grupos', 'significativos', 'de', 'ciudadanos', 'y', 'movimient
os', 'sociales', 'vence', 'el', '13', 'de', 'noviembre', 'del', '2021', ',', 'esto', ',', 'de', 'acuerdo', 'con', 'lo', 'establ
```

Posteriormente se consultan cuales son las stopwords en el idioma que queramos, en este caso es el español, y se imprimen, paso solamente es para saber cuales palabras se ignorarán en el proceso de análisis.

Stopwords

Stop words are basically a set of commonly used words in any language, not just English.

The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead.

```
stop_words = set(stopwords.words("spanish"))
print(stop_words)
```

```
{'estéis', 'fueron', 'algunos', 'han', 'habrá', 'habías', 'estamos', 'e', 'hube', 'mío', 'fuisteis', 'esa', 'estuviesen', 'estad', 'donde', 'seríais', 'él', 'tanto', 'será', 'estarían', 'a', 'por', 'fueses', 'tengáis', 'estados', 'esas', 'eres', 'estoy', 'tendría', 'el', 'tuvo', 'ni', 'erais', 'hubiesen', 'tuyo', 'es', 'tendrían', 'serán', 'otras', 'seré', 'tengo', 'hubo', 'tuvieran', 'vosotras', 'muy', 'sintiendo', 'cual', 'sois', 'fueras', 'estuvieseis', 'quien', 'estés', 'nuestros', 'tuve', 'suyo', 'contra', 'estuvierais', 'estás', 'te', 'para', 'nada', 'tenía', 'tuvimos', 'ella', 'sin', 'mucho', 'fue', 'estar', 'fuese', 'tuviéramos', 'otros', 'tenías', 'como', 'habréis', 'hubiera', 'otro', 'me', 'tuvisteis', 'pero', 'sean', 'tendrías', 'sería', 'hubimos', 'su', 'habíamos', 'habéis', 'estaré', 'hayáis', 'habrías', 'éramos', 'estáis', 'o', 'sentida', 'habríais', 'serías', 'estada', 'estaréis', 'hayamos', 'fui', 'tendremos', 'estuvieron', 'esto', 'y', 'poco', 'suyos', 'tenida', 'serás', 'tenidos', 'míos', 'tuviera', 'las', 'al', 'sobre', 'nuestra', 'los', 'hubisteis', 'soy', 'fuerais', 'vuestros', 'fuéramos', 'nosotros', 'estuviste', 'hubieras', 'la', 'nuestro', 'todo', 'sentido', 'una', 'ellos', 'tendréis', 'tenéis', 'tuviese', 'habidas', 'hemos', 'yo', 'tendré', 'unos', 'tuviesen', 'tendrás', 'tenido', 'ellas', 'estuviésemos', 'tiene', 'tus', 'serían', 'fueseis', 'siente', 'habido', 'sus', 'nos', 'tenían', 'durante', 'estuviéramos', 'mía', 'son', 'hasta', 'estos', 'estén', 'del', 'habremos', 'hubiéramos', 'cuando', 'estadas', 'sí', 'eso', 'mí', 'tuyas', 'vuestro', 'seremos', 'ante', 'habríamos', 'estabais', 'hubieseis', 'ti', 'mías', 'haya', 'esta', 'con', 'tengamos', 'estuviese', 'fuera', 'he', 'teníais', 'se', 'qué', 'estarías', 'tuyos', 'estarás', 'habrán', 'uno', 'seamos', 'mis', 'ha', 'ya', 'sentidos', 'estuviera', 'este', 'fueran', 'tuvieron', 'fuimos', 'habiendo', 'seas', 'tuya', 'tengas', 'tienen', 'estuvieras', 'antes', 'estuvieran', 'algunas', 'estaríamos', 'esté', 'nosotras', 'vuestra', 'muchos', 'hayas', 'habré', 'están', 'mi', 'tenemos', 'quienes', 'estaremos', 'hubieran', 'tendrán', 'estuvimos', 'también', 'estarán', 'desde', 'hubiste', 'tenidas', 'tuvierais', 'os', 'no', 'somos', 'lo', 'eran', 'en', 'ese', 'habrían', 'tuviese', 'estaríais', 'tu', 'estuve', 'fuesen', 'estuvo', 'vosotros', 'sentidas', 'tendríais', 'hubieses', 'has', 'estuvieses', 'habrás', 'seríamos', 'tuvieras', 'más', 'que', 'habida', 'estabas', 'teniendo', 'seáis', 'estaban', 'fuésemos', 'había', 'fuiste', 'habidos', 'tenga', 'esos', 'les', 'de', 'le', 'está', 'tuviste', 'tú', 'estaría', 'era', 'sentid', 'tuvieseis', 'suya', 'estando', 'habían', 'nuestras', 'estaba', 'hayan', 'estábamos', 'todos', 'entre', 'eras', 'estado', 'estas', 'porque', 'estuvisteis', 'hubierais', 'algo', 'teníamos', 'tuviésemos', 'hubiésemos', 'hubiese', 'tendríamos', 'sea', 'estará', 'tengan', 'estemos', 'hubieron', 'otra', 'habría', 'tendrá', 'habíais', 'vuestras', 'hay', 'tienes', 'tened', 'suyas', 'un', 'seréis'}
```

Luego imprimimos las palabras filtradas sin las stopwords.

```
'temeroso', 'posible', 'incumplimiento', 'acuerdo', 'parte', 'gobierno', ',', 'londoño', 'pidió', 'ayuda', 'comunidad', 'internacional', 'dar', 'información', 'situación', 'jurídica', 'compañero', ',', 'salió', 'país', 'permiso', 'jep', 'evento', 'político', 'ciudad', 'méxico', '.', 'corte', 'suprema', 'amenaza', 'quedarse', 'pocas', 'mujeres', 'magistradas', 'sala', 'plena', 'magistradas', 'hilda', 'gonzález', 'patricia', 'salazar', 'dos', 'únicas', 'mujeres', 'hacen', 'parte', 'sala', 'plena', 'corte', 'suprema', 'justicia', ',', 'momento', ',', 'cuota', 'amenaza', 'desaparecer', 'vez', 'culmine', 'periodo', 'última', 'abril', 'próximo', ',', 'si', 'llenan', 'plazas', 'cinco', 'dignatarios', 'alto', 'tribunal', '.', 'aunque', 'ley', 'cuotas', 'prometen', 'cumplirse', 'entidades', 'públicas', ',', 'poder', 'parece', 'ser', 'alcanzado', 'corte', 'suprema', 'justicia', '.', 'sala', 'plena', 'alto', 'tribunal', ',', 'suele', 'tener', 'nómina', '23', 'magistrados', ',', 'momento', ',', 'cuenta', 'solo', '19', 'debido', 'restantes', 'cumplido', 'periodo', '.', 'llama', 'atención', 'ocupan', 'actualmente', 'plazas', 'solo', 'dos', 'mujeres', ',', 'hilda', 'gonzález', 'patricia', 'salazar', ',', 'última', 'abril', 'próximo', ',', 'tras', 'ocho', 'años', 'labor', ',', 'deberá', 'abandonar', 'cargo', '.', 'dilema', 'ley', 'cuotas', ',', 'ocasión', ',', 'iniciativa', 'norma', 'establece', 'debe', 'garantizar', '30', '%', 'participación', 'mujer', 'tres', 'ramas', 'poder', 'público', 'entidades', 'carácter', 'nacional', ',', 'departamental', ',', 'regional', ',', 'provincial', ',', 'distrital', 'municipal', '.', 'orden', 'ideas', ',', 'si', 'sala', 'plena', 'corte', 'suprema', 'acogiera', 'cumpliera', 'designio', 'debería', 'tener', 'nómina', 'menos', 'seis', 'magistradas', 'total', '23', 'juristas', 'componen', '.', 'conocer', 'justicia', ',', 'seguridad', 'derechos', 'humanos', ',', 'visite', 'sección', 'judicial', 'espectador', '.', 'aunque', 'exmagistrada', 'fanny', 'gonzález', ',', 'falleció', 'toma', 'retoma', 'palacio', 'justicia', '1985', ',', 'abrió', 'puerta', 'mujeres', 'llegaran', 'ocupar', 'altos', 'rangos', 'justicia', 'colombiana', ',', 'participación', 'género', 'suele', 'ser', 'escasa', ',', 'pues', 'tras', 'muerte', ',', 'solo', '16', 'años', 'después', ',', 'mujer', 'condición', 'magistrada', 'volvió', 'recorrer', 'pasillos', 'alto', 'tribunal', '.', 'investigación', 'dejusticia', 'muestra', 'composición', 'corte', 'suprema', 'reveló', 'panorama', 'alentador', 'presentó', '2012', '2014', 'alcanzó', ',', 'promedio', ',', '20', '%', 'representación', 'mujer', 'corporación', '80', '%', 'hombres', '.', 'investigación', 'da', 'cuenta', '2015', '2017', 'cifra', 'descendió', '2', '%', 'alcanzando', '18', '%', 'participación', 'mujeres', 'alto', 'tribunal', ',', 'posicionando', 'estadística', 'punto', 'favorecedor', 'género', 'masculino', 'llegando', '82', '%', '.', 'pese', ',', 'alguna', 'forma', ',', 'máximo', 'tribunal', 'penal', 'hace', 'años', 'determinó', 'acogerían', 'propios', 'lineamientos', 'escoger', 'dignatarios', ',', 'si', 'cierto', 'consejo', 'superior', 'judicatura', 'cumplido', 'denominada', 'cuota', 'mujeres', 'listas', '.', 'ejemplo', ',', 'reemplazar', 'exmagistrado', 'luis', 'armando', 'tolosa', ',', 'encuentra', 'dignatario', 'corte', ',', 'corporación', 'envió', 'lista', 'nombre', 'seis', 'mujeres', 'cuatro', 'hombres', '.', 'lea', 'aquí', ':', 'karen', 'abudinen', 'llamada', 'declarar', 'corte', 'suprema', 'caso', 'mintic', 'reemplazar', 'exmagistrada', 'clar
```

Se realiza un histograma de las palabras más frecuentes de nuestro análisis (ignorando los signos de puntuación).

Dichas palabras son:

1. Años.
2. Colombia.
3. Puede.
4. Ser.
5. Salud.
6. Si.
7. Personas.

[12]:
```
frec_dist = FreqDist(filtered_words)

print(frec_dist)
```

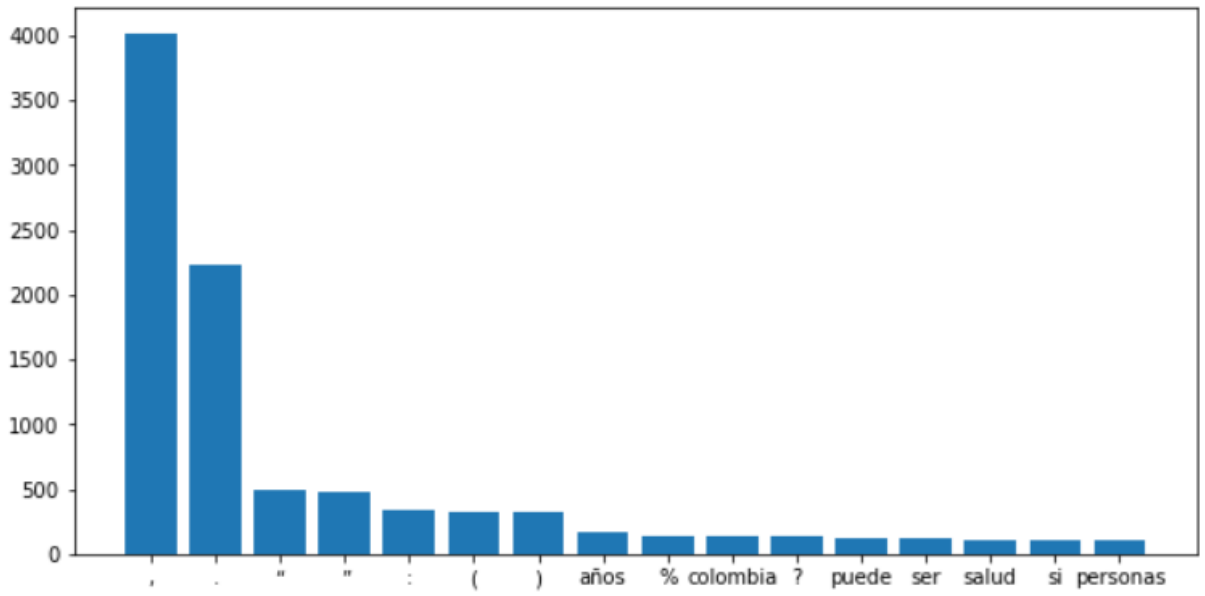
```
<FreqDist with 9454 samples and 43650 outcomes>
```

[13]:
```
frec_dist.most_common(5)
```

[13]:
```
[(',', 4015), ('.', 2232), ('“', 489), ('”', 482), (':', 331)]
```

[19]:
```
n = 16
plt.figure(figsize=(10,5))
plt.bar([ w[0] for w in frec_dist.most_common(n) ], [ w[1] for w in frec_dist.most_common(n) ])
```

[19]:
```
<BarContainer object of 16 artists>
```



Y su respectiva nube de palabras, fundamental para interpretar los resultados de nuestro análisis.

```
[20]: wordcloud = WordCloud(max_font_size = 50, max_words = 100, background_color = "white").generate(" ".join(filtered_words))

      plt.imshow(wordcloud, interpolation = "bilinear")
      plt.axis("off")
      plt.show()
```

Para extender mejor el taller se hizo un cambio de las palabras más frecuentes por término de búsqueda.

```
[4]: list(collection.find({'$text': {'$search': 'covid'}}, {'_id': 0, 'summary': 1, 'title': 1, 'full_text': 1,}))
```

```
[4]: [{'title': 'Combinar vacunas AstraZeneca, Pfizer y Moderna es efectivo contra el Covid-19 ',
   'summary': 'La combinación de la vacuna AstraZeneca con las de ARNm como Pfizer y Moderna aumentan la efectividad contra el Covid-19, según dice un estudio sueco recién publicado.',
   'full_text': 'Las personas que han recibido una primera dosis de la vacuna contra la covid-19 de Oxford/AstraZeneca y una segunda de ARNm tenían menor riesgo de infección frente a las que reci
bieron ambas dosis del primer preparado, según un estudio de investigadores de la Universidad de Umea (Suecia). En las vacunas de ARN mensajero (ARNm), como la de Pfizer-BioNTech y Moderna, se e
mplea ácido ribonucleico para lograr el desarrollo de una respuesta inmune. El profesor de medicina geriátrica de la citada universidad, Peter Nordström, que ha realizado la investigación, ha re
saltado que cualquier vacuna aprobada es mejor que ninguna. (Para más información confiable sobre vacunas y coronavirus, lea El Espectador) “Sin embargo, nuestro estudio muestra una mayor reducc
ión del riesgo para las personas que recibieron una vacuna de ARNm después de haber recibido una primera dosis de una vacuna basada en vectores, en comparación con las personas que recibieron la
vacuna basada en vectores para ambas dosis”, añadió el especialista, cuyo estudio ha publicado la revista Lancet Regional Health. (En estos casos Colombia permitirá combinar terceras dosis con P
fizer y Moderna) Al detenerse el uso de la vacuna basada en vectores de Oxford-AstraZeneca para personas menores de 65 años, a todas las personas que ya habían recibido su primera dosis de este
preparado se les recomendó una de ARNm como segunda dosis. Durante un período de seguimiento promedio de más de dos meses después de la segunda dosis, el estudio mostró un 67% menos de riesgo de
infección para la combinación de Oxford-AstraZeneca y Pfizer-BioNTech, y un 79% menos de riesgo para Oxford/AstraZeneca y Moderna, ambos en comparación con los individuos no vacunados. Para las
personas que recibieron dos dosis de la vacuna Oxford-AstraZeneca, la reducción del riesgo fue del 50%. Estas estimaciones de riesgo se observaron después de tener en cuenta las diferencias con
respecto a la fecha de vacunación, la edad de los participantes, el estado socioeconómico y otros factores de riesgo de covid-19. Los investigadores han puntualizado que las estimaciones de efic
acia se aplicaron en relación con la infección por la variante Delta, que dominaba los casos durante el período de seguimiento. Investigaciones anteriores han demostrado que los programas de vac
unación combinados generan una sólida respuesta inmunitaria, pero no está claro hasta qué punto estos esquemas pueden reducir el riesgo de infección clínica, según el análisis. El estudio de Ume
a, en el que se analizaron los casos de 700.000 personas, se basa en datos de registros nacionales de la Agencia de Salud Pública de Suecia, la Junta Nacional de Salud y Bienestar y Estadísticas
de Suecia. En Colombia es permitida la posibilidad de combinar vacunas de diferentes farmacéuticas desde el pasado 7 de septiembre, y ya hay lineamientos técnicos y operativos de la aplicación d
e las vacunas contra el COVID-19, entre esos, los casos en los que se podrá combinar la vacunación inicial con terceras dosis de las farmacéuticas Pfizer y Moderna. Según el documento, que tiene
fecha del pasado 7 de septiembre, para la aplicación de una tercera dosis de refuerzo contra el COVID-19 se podrá implementar la “intercambiabilidad de vacunas” solo en personas con inmunosupres
ión y siempre y cuando la vacuna inicial haya sido de plataforma de vector viral. Es decir, que el esquema de vacunación inicial haya sido con los biológicos de Janssen o de AstraZeneca. Si es u
na persona con inmunosupresión y su esquema inicial de vacunación fue de virus inactivado o de ARNm (como las vacunas de Pfizer y Moderna) la dosis de refuerzo tendrá que ser con un biológico ho
mólogo.'},
  {'title': 'Científicos han recibido amenazas y ataques por hablar del COVID-19 en los medios',
   'summary': 'Una encuesta realizada por la revista científica Nature a más de 300 investigadores que han concedido entrevistas a medios de comunicación sobre el COVID-19 encontró una amplia exp
eriencia de acoso, abuso, amenazas de violencia física y sexual. El 15% dijo haber recibido amenazas de muerte.',
   'full_text': 'Numerosos científicos han manifestado públicamente que han sufrido acoso tras hablar del coronavirus en pandemia. La revista Nature\xa0analiza el problema en un trabajo que ha ll
evado a cabo a partir de una encuesta a 321 científicos que han hablado sobre el COVID-19 con los medios o en redes sociales. Un 15 % de las personas encuestadas han recibido amenazas de muerte
y un 22 %, amenazas de violencia física o sexual. (Le recomendamos: ¿Cuándo dejaremos de usar tapabocas en Colombia? No será en el partido contra Ecuador) Nature\xa0relata, por ejemplo, el caso
de la médica de enfermedades infecciosas\xa0Krutika Kuppalli, que llevaba apenas una semana en su nuevo trabajo en septiembre de 2020, cuando alguien la llamó por teléfono a su casa y la amenazó
de muerte. Kuppalli, que acababa de mudarse de California (EE. UU.) a la Universidad Médica de\xa0Carolina del Sur, en Charleston, llevaba meses lidiando con el acoso\xa0online\xa0tras haber dad
o entrevistas en los medios de comunicación sobre el COVID-19 y su miedo aumentó al recibir una llamada amenazándola de muerte. “Me puso muy ansiosa, nerviosa y alterada”, declara Kuppalli, que
ahora trabaja en la Organización Mundial de la Salud (OMS) en Ginebra, Suiza.\xa0Ante una segunda llamada parecida la policía le sugirió que se comprase un arma. La experiencia de Kuppalli duran
te la pandemia no es aislada. (Puede leer más sobre coronavirus y salud aquí) El trabajo de\xa0Nature, que se basa en un primer sondeo que realizó el\xa0Science Media Centre australiano\xa0(AusS
MC) entre investigadores de COVID-19, indica que la situación es preocupante. Del SMC de Australia recogieron un total de\xa050 testimonios, de los cuales seis (12 %) declararon haber recibido a
menazas de muerte y otros seis dijeron haber recibido amenazas de violencia física o sexual. La encuesta recoge agresiones que van desde ataques a la credibilidad y a la reputación hasta amenaza
s físicas, que en algunos casos se materializaron. “Seis científicos dijeron que habían sido atacados físicamente. La encuesta se realizó de forma anónima, pero sabemos que su procedencia era: t
res en el Reino Unido, uno en Nueva Zelanda y dos en EE UU. También ha habido algunas denuncias públicas de agresiones, como la del consejero científico jefe del Reino Unido,\xa0Chris Whitty, qu
e fue agarrado y empujado en la calle. Algunos científicos denunciaron también el envío de cartas con pólvora a su casa”, dice a SINC,\xa0Richard Van Noorden, editor de la revista. (Le puede int
eresar: ¿Tomar una aspirina diaria por prevención? Científicos ya no lo recomiendan) La encuesta se distribuyó por correo electrónico después a científicos que contribuyen habitualmente en los c
ontenidos de los\xa0SMC del Reino Unido, Alemania, Canadá, Taiwán y Nueva Zelanda.\xa0Nature\xa0también la envió a investigadores de EE. UU. y Brasil que habían sido citados de forma destacada e
```

Link del video: https://youtu.be/6EQCEEm358Q