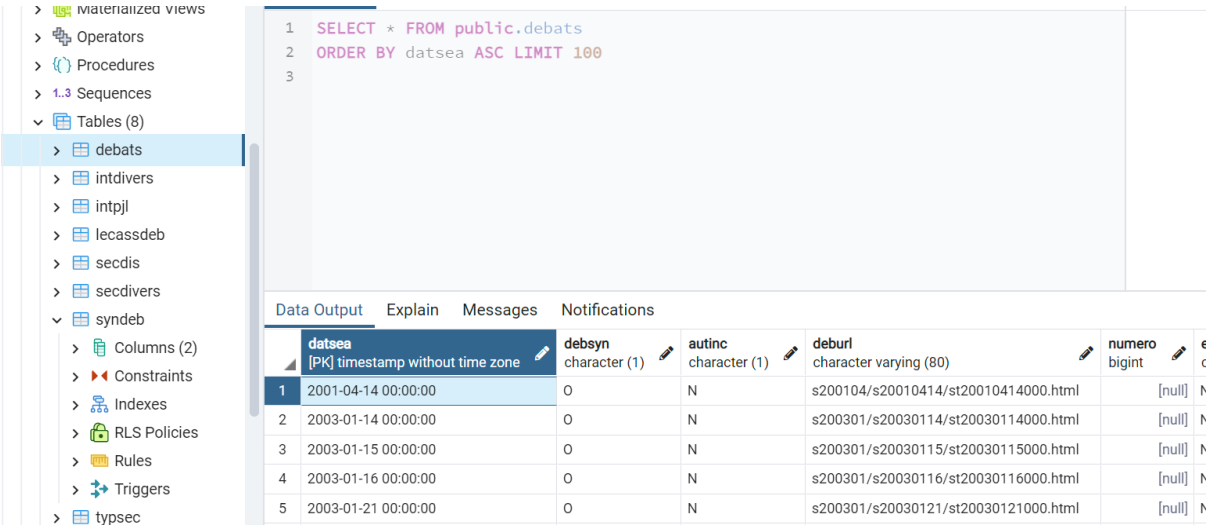
Parse du Sénat - *Parse french senate*

Data

All the debate URLs

Source and install

Data source: https://data.senat.fr/la-base-comptes-rendus/



Installed PostgreSQL on Windows, imported dataset thanks to pgAdmin4 console. Used the psql console, then import command:

$ \i 'c:/Users/myname/some path/query.sql'

/!\ Respect the simple quotes, and forward slashes. Spaces can be included safely.

Example

First line in the database, table “debats”: index: “2001-04-14 00:00:00”: column “deburl” : s200104/s20010414/st20010414000.html

When we try the “/seances” route, we get the complete debate.

https://www.senat.fr/seances/s200104/s20010414/st20010414000.html

We now have a list of all URLs to crawl if we need to have all the debates.

Single debate page

Source

The route is “/seances”, and the URI comes from the previous section:
https://www.senat.fr/seances/s200104/s20010414/st20010414000.html

Screenshot not shown here, as it is as long as a 22 page PDF file. The main features are anchor points inside the documents, with summary sections acting as duplicate information and context, that is useless, on top of the actual debate.
We might filter the useless data out, or not, depending on the resulting extracted words (if frequency filters are useless, then we might filter on HTML tags).

Example 1

From the useless page to the useful one, we remove the t and increment the last number :

https://www.senat.fr/seances/s200104/s20010414/st20010414000.html
https://www.senat.fr/seances/s200104/s20010414/s20010414001.html

Example 2

If we remove the “t” from the last URI and increment by one, we have the actual URL:
https://www.senat.fr/seances/s200301/s20030116/st20030116000.html
https://www.senat.fr/seances/s200301/s20030116/s20030116001.html

First colab : download all 2500 debates

Link to colab

Colab is here :
https://colab.research.google.com/drive/1WmQbINvbIFN4m0SrlXX-A4irqqhf8VXP#scrollTo=ehCSzEO229gm

And related dataset is here :
https://drive.google.com/drive/folders/1nu3o0kQysQsayQRLhmPZnaQaH3epvnNI

Steps taken

1. extract list of raw URI to edbate summary page
2. convert to full summary page (replace “st” by “s”, replace “000” by “001”)
3. generate list of complete URLs (domain + complete URI)
4. multithreaded functions to scrape target URLs
5. slugify used to sanitize URLs to filenames
6. save all as txt files

We now have a list of 2.5k+ txt files containing the HTML pages of each debate.

Second colab : build dataset

We now need to build the dataset from the raw HTML text files. To do this, we use BeautifulSoup to parse html and Pandas to store everything in dataframes and export to csv.

Link to colab

https://colab.research.google.com/drive/1Z-EWhGR5XKjtMhpxx9A5PYvEy3NmqBtA#scrollTo=A87CUQ1GHVoC
And related dataset is here (same place as first notebook) :
https://drive.google.com/drive/folders/1nu3o0kQysQsayQRLhmPZnaQaH3epvnNI

Steps taken

1. lib import and data from previous step (2.5k text files) are loaded
2. single parse of a debate
3. parse of all debates
4. concat and save to csv file

We now have a 110k samples, with the following features:

- date : date of the whole debate
- title : title of the whole debate
- speaker_name : name of the current intervention speaker
- speaker_quality : eg. minister of, mostly empty
- speaker_link : href to profile of speaker
- speaker_intervention : the actual intervention content

A debate is a collection of interventions.
Each intervention contains a text, and is made by a speaker.
The speaker has a name, quality, and profile link.
Date and title are meta and apply to all interventions of a given debate.

Third colab : General stats

We check the dates on calendars and compare file sizes to check if the dataset is mostly complete or mostly wrong.

This notebook is actually split in many parts due to the size of each sub-section.

Part A : Sex extraction and women equality

Link to colab

https://colab.research.google.com/drive/19GjYTf2dFQ5WqhHsR4H4nkUq6txi-FL2#scrollTo=Na8UwaPHEQa1

Steps taken

1. added sex feature by parsing “M.” or “Mme” from the “speaker_name” feature
2. displayed bar plots for the weekdays for :
    a. dataset with all interventions
    b. dataset without director
    c. participation mode only: any number of interventions in a debate count for 1
3. compared the above 3 analysis in a line plot
4. displayed women parity rations on calendars from years 2001 to 2022
5. analyzed men and women attendance number repartition to find relevant scale values
6. display women attendance with said custom scale
7. TODO: compare women interventions from one generation to the other (one generation is the samples that exist between two elections)

What we learned is that women are discriminated against, more or less depending how we measure the phenomenon, but they definitely are being silenced. From the attendance numbers, we know that they are present, but from the intervention count, they do not speak often, or lack the back-and-forth attitude that makes the director and some men speakers so prevalent in the dataset.

On the weekday analysis, the 3 middle-days (tuesday, wednesday, and thursday) are the most important days in terms of attendance and intervention numbers. They also show a high discrimination rate as many more men are present these days, while absolute women numbers do not move much (doubling a small number of women is negligeable compared to a 25% increase in a large number of men).

Part B : Nobility extraction and representativity

We will look out for nobility given names as they are social clues. We will ask ourselves whether the person holding the title is legitimate, where do they come from, how many are there compared to the general population.

Link to colab

https://colab.research.google.com/drive/15r4hLkxnfzTjRpMP9Y0_NXEQBqsqLYeM#scrollTo=WdFFVOuNHizV