

# Parse du Sénat - *Parse french senate*

<b>Data</b>	<b>1</b>
All the debate URLs	1
Source and install	1
Example	2
Single debate page	3
Source	3
Example 1	3
Example 2	3
<b>First colab : download all 2500 debates</b>	<b>3</b>
Link to colab	3
Steps taken	4
<b>Second colab : build dataset</b>	<b>4</b>
Link to colab	4
Steps taken	4
<b>Third colab : General stats</b>	<b>5</b>
Part A : Sex extraction and women equality	5
Part B : Nobility extraction and representativity	5

## Data

### All the debate URLs

#### Source and install

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

- Materialized views
- Operators
- Procedures
- Sequences
- Tables (8)
  - debats
  - intdivers
  - intpjl
  - lecassdeb
  - secdis
  - secdivers
  - syndeb
    - Columns (2)
    - Constraints
    - Indexes
    - RLS Policies
    - Rules
    - Triggers
  - typsec

```

1 SELECT * FROM public.debats
2 ORDER BY datsea ASC LIMIT 100
3

```

Data Output

	datsea [PK] timestamp without time zone	debsyn character (1)	autinc character (1)	debur character varying (80)	numero bigint	
1	2001-04-14 00:00:00	O	N	s200104/s20010414/st20010414000.html	[null]	
2	2003-01-14 00:00:00	O	N	s200301/s20030114/st20030114000.html	[null]	
3	2003-01-15 00:00:00	O	N	s200301/s20030115/st20030115000.html	[null]	
4	2003-01-16 00:00:00	O	N	s200301/s20030116/st20030116000.html	[null]	
5	2003-01-21 00:00:00	O	N	s200301/s20030121/st20030121000.html	[null]	

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 “debur”: s200104/s20010414/st20010414000.html

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



**Bienvenue au Sénat**  
*Un site au service des citoyens*

Vous êtes ici : Travaux parlementaires > Comptes rendus > Comptes rendus intégraux de avril 2001

**SÉANCE DU 14 avril 2001 (compte rendu intégral des débats du Sénat)**

**SOMMAIRE**

**PRÉSIDENCE DE M. ADRIEN GOUTEYRON**

**1. Procès-verbal**

**2. Communications électroniques. - Suite de la discussion d'un projet de loi déclaré d'urgence**

[Article 11](#)

Amendement n° 279 de M. René Trégouët. - MM. René Trégouët, Pierre Hérisson, rapporteur de la commission des affaires économiques ; Patrick Devedjian, ministre délégué à l'Industrie. - Retrait.

Amendement n° 256 de M. René Trégouët. - MM. René Trégouët, Pierre Hérisson, rapporteur ; le ministre délégué. - Retrait.

Adoption de l'article.

[Article 12. - Adoption](#)

[Articles additionnels avant l'article 13](#)

Amendement n° 204 de Mme Marie-France Beaufils. - Mme Odette Terrade, MM. Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

Amendement n° 210 de Mme Marie-France Beaufils. - MM. Gérard Le Cam, Pierre Hérisson, rapporteur ; le ministre délégué, Mme Marie-France Beaufils. - Rejet.

Amendement n° 207 de Mme Marie-France Beaufils. - Mme Odette Terrade, MM. Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

Amendement n° 211 de Mme Marie-France Beaufils. - Mme Marie-France Beaufils, MM. Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

Amendement n° 209 de Mme Marie-France Beaufils. - Mme Odette Terrade, MM. Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

Amendement n° 208 de Mme Marie-France Beaufils. - MM. Gérard Le Cam, Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

Amendement n° 205 de Mme Marie-France Beaufils. - Mme Marie-France Beaufils, MM. Pierre Hérisson, rapporteur ; le ministre délégué. - Rejet.

<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 ratios 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 negligible 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](https://colab.research.google.com/drive/15r4hLkxnfzTjRpMP9Y0_NXEQBqsqLYeM#scrollTo=WdFFVOuNHizV)