Web Scraping and Social Media Scraping project desctiption

May 9, 2021

- Author: Wojciech Maślakiewicz
  - Student ID: 394350

---

1 Description

1.0.1 Motivation and webpage

Unbiasednes, transparency and data-based facts are a rare phenomenon in media coverage on Polish politics. To make a small contribution to changing this state I scraped all the available voting sessions of the polish Senate to obtain the individual votes of every senator in every voting for which the results are available on the Senate's webpage. The data was collected from the official Senate webpage, where the voting outcomes are available starting from the year 2015 (VIII term) up to now. With such data one can easily perform many useful and interesting analsis providing data-based insights into the actual voting behaviours of the MPs. Next section briefly describes the scraping process.

1.0.2 Scraping mechanics and the data collected

The scraping had three main stages: 1. *Collecting information about the sessions:*

Information about the title, dates and votings that took place on a given day was obtain from subpages of the main age. See example link

For every session the scrapers collect: * the title of the session (in the example the text **Posiedzenie: 21. posiedzenie Senatu RP X kadencji, 2 dzień**)

- the dates of the session (in the example the text **17, 18 i 19 lutego 2021 r.**)
- information about each voting that took place that day:
  - number of the voting (**3,4,5,6**)
  - title of the voting (**Głosowanie próbne, ... , Drugie czytanie projektu uchwały w 50. rocznicę strajku włókniarek w Łodzi**)
  - subscript of the title if available (**Wniosek o przyjęcie projektu**)
  - link to the voting results for each senator

2. *Collecting information about each voting:*

For every voting link obtained from step 1. the individual votes of every senator are retrieved. Precisely, every link like this example link produces a table with 100 rows (one for each senator) and 3 columns (name, vote, voting_id).

3. *Collecting the party affiliations of each MP in a given term of office:* From the page (three such pages, as three terms (8,9,10) are considered) the membership of each MP to the club/party is obtained.

Stages 1. and 2 are dependent i.e the second stage is 'fed' by the links obtained in the first stage. The 3. stage is independent from the two previous ones and can be run separately. The total runtime in the last section is the runtime of all the three stages added together.

1.0.3 The obtained data

This section provides technical description of the data and example of 5 first lines of the tables obtained from scraping.

Sessions *Columns (7)*

- **dates:** the date(s) in which the session took place
- **session_title:** the title of the sesison containing its number and the number of Senate's term of office
- **voting_n0:** the number of voting in the session (eg. 1 is the first voting in the session)
- **voting title:** the title containing the subject of the voting
- **subscript:** (for some votings) a subtitle specifying the nature of the voting
- **link:** link to the voting results
- **voting_id:** ID to join with other tables

Rows (9991)

Each row corresponds to one voting.

	dates	link	session_title	subscript	voting_id	voting_no	voting_title
0	17 listopada 2011 r.	link	Posiedzenie: 2. posiedzenie Senatu RP VIII kadencji, 1 dzień	Wniosek o podjęcie uchwały	25	1	Powołanie Komisji Regulaminowej, Etyki i Spraw Senatorskich.
1	17 listopada 2011 r.	link	Posiedzenie: 2. posiedzenie Senatu RP VIII kadencji, 1 dzień	Wniosek o podjęcie uchwały	24	2	Powołanie stałych komisji senackich.

	dates	link	session_title	subscript	voting_id	voting_no	voting_title
2	20, 21 i 22 grudnia 2011 r.	link	Posiedzenie: 3. posiedzenie Senatu RP VIII kadencji, 1 dzień	Wniosek o przyjęcie ustawy bez poprawek	2	1	Ustawa o zmianie niektórych ustaw związanych z realizacją ustawy budżetowej
3	20, 21 i 22 grudnia 2011 r.	link	Posiedzenie: 3. posiedzenie Senatu RP VIII kadencji, 1 dzień	Poprawka 1	1	2	Ustawa o zmianie niektórych ustaw związanych z realizacją ustawy budżetowej
4	20, 21 i 22 grudnia 2011 r.	link	Posiedzenie: 3. posiedzenie Senatu RP VIII kadencji, 1 dzień	Poprawka 2, 5	4	3	Ustawa o zmianie niektórych ustaw związanych z realizacją ustawy budżetowej

Votes *Columns (3)*
* **name:** name of the MP
* **vote:** the vote casted {'za', 'nieob.', 'przec.', 'wstrz.', 'nie gł.'}
* **votind_id:** ID to join with other tables

Rows (994725)

Each row corresponds to one vote casted of one senator in one voting.

	name	vote	voting_id
0	Ł.M. Abgarowicz	za	25
1	Ł.M. Abgarowicz	za	22
2	Ł.M. Abgarowicz	za	17
3	A.T. Aksamit	za	25
4	T. Arłukowicz	za	25

Clubs *Columns (3)*
* **club:** name of the club or party
* **name:** name of the MP
* **term:** number of the Senate's term of office and an ID to join with the sessions table (*)

Rows (342)

Each row corresponds to one senator in a given term of office.

(*) actual join can be done after some preprocessing of the sessions table which is out of the scope of this project

	club	name	term
0	Klub Parlamentarny Prawo i Sprawiedliwość	Rafał Ambrozik	9
1	Klub Parlamentarny Prawo i Sprawiedliwość	Grzegorz Bierecki	9
2	Klub Parlamentarny Prawo i Sprawiedliwość	Przemysław Błaszczyk	9
3	Klub Parlamentarny Prawo i Sprawiedliwość	Aleksander Bobko	9
4	Klub Parlamentarny Prawo i Sprawiedliwość	Margareta Budner	9

2 Analysis

There are many interesting analysis that can be done with this kind data. For an example of more extended analysis see my unsupervised learning final project where I analysed analogical data for Sejm. This data, after very little preprocessing, can be used to perform an analogical analysis for the other chamber of polish parliment

Nonetheless, to comply with the requirements for this project let's perform some symplistic analysis.

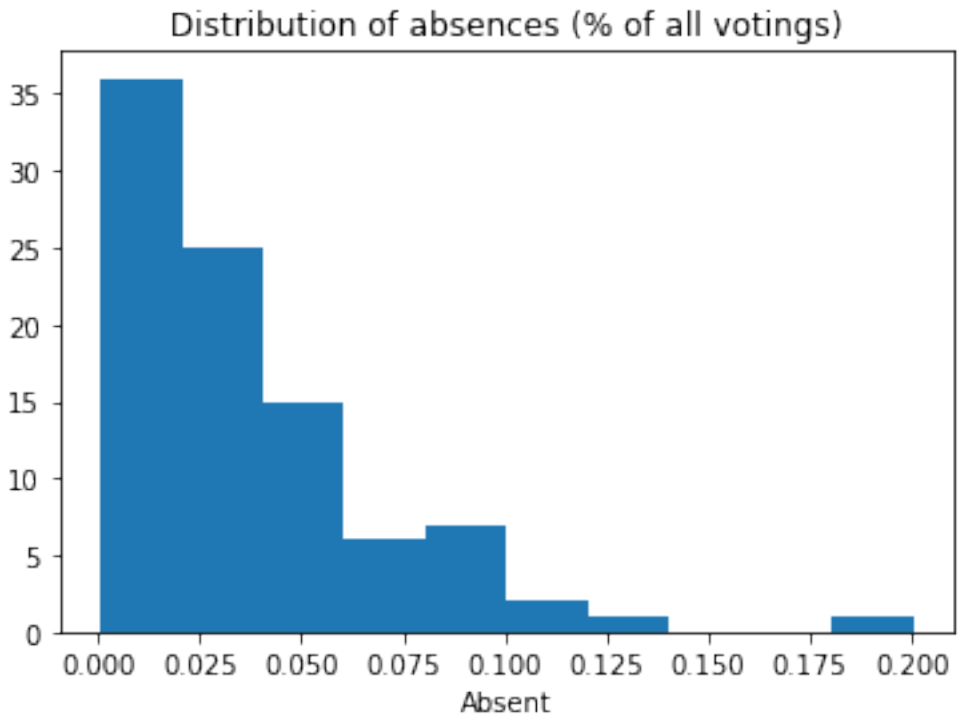
	name	For
0	J.A. Wyrowiński	0.764811
1	L. Czarnobaj	0.721065
2	J. Michalski	0.716207
3	Z.H. Meres	0.715983
4	H. Górski	0.708333

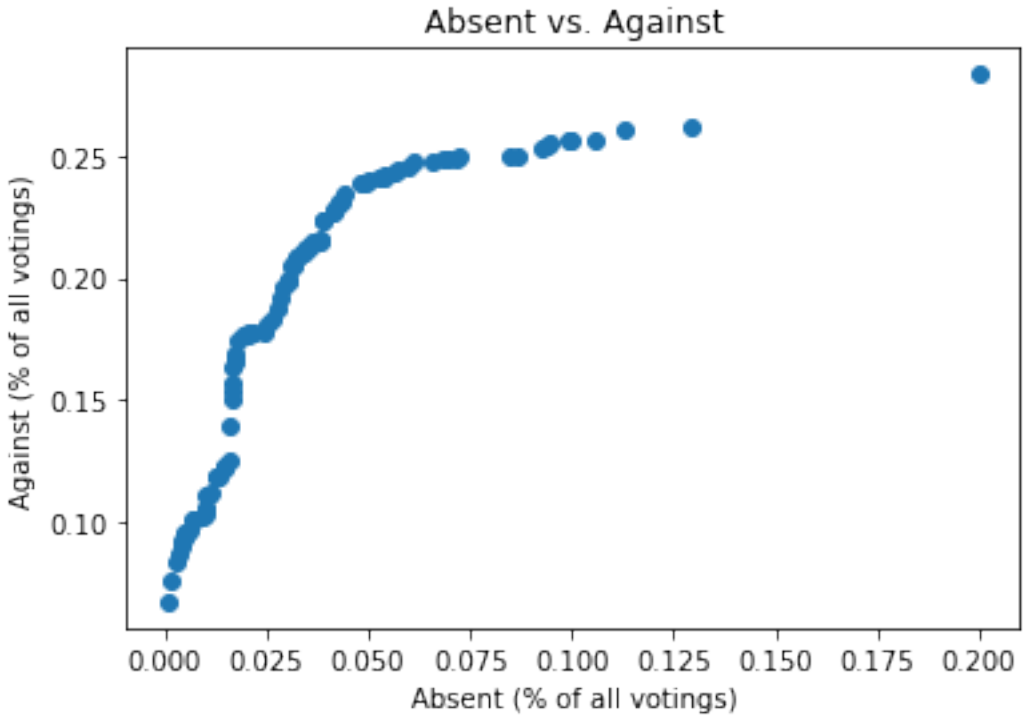
	name	Against
0	J. Wyrowiński	0.283513
1	Z.H. Meres	0.261601
2	J. Michalski	0.260704
3	P.A. Gruszczyński	0.256221
4	J.M. Sepioł	0.255996

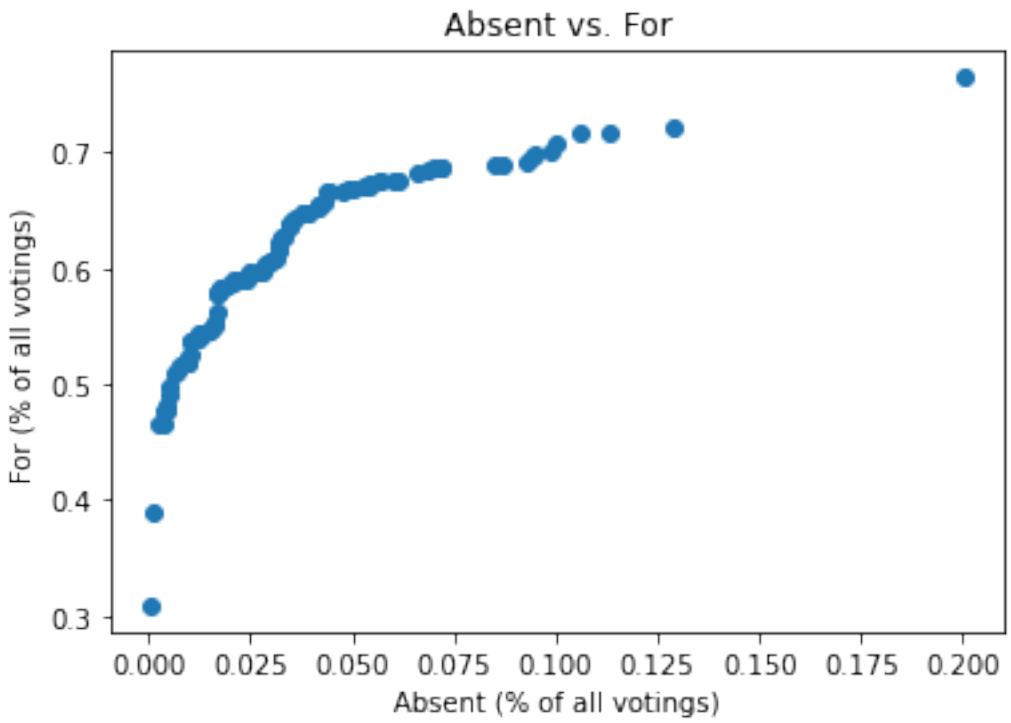
	name	Absent
0	W.Z. Ortyl	0.200157
1	A.T. Aksamit	0.129119
2	T. Arłukowicz	0.112979
3	A. Szewiński	0.10603
4	B.M. Pęk	0.0999776

We may see which MPs most frequently (as percentage of total votes casted) voted For, Against or were absent :

We may even look at some scatterplots:


Distribution of absences (% of all votings)
35
30
25
20
15
10
5
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Absent


Absent vs. Against
Against (% of all votings)
0.25
0.20
0.15
0.10
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Absent (% of all votings)



We may produce hundreads of simmilar plots for different variables and aggregation levels (grouping by parites, votings, sessions, dates etc.)

3 Comparison of runtime

Without surprise, scrapy was the fastest and selenium was the slowest.

place	framework	total runtime (sec)	total runtime (hr)	compared to best (ratio)
1	Scrapy	1146.98	0.32	1
2	Beautiful Soup	7021.64	1.95	6.1
3	Selenium	41745.23	11.6	37.4