Fakulta matematiky, fyziky a informatiky Univerzity Komenského, Bratislava

Predikcie volieb podľa prieskumov verejnej mienky

projekt z predmetu Princípy dátovej vedy

Tomáš Antal
Erik Božík
Teo Pazera
Andrej Špitalský
Tomáš Varga

22. decembra 2024

Obsah

1 Predstavenie témy

2 Spracovanie dát

2.1 Scraping dát

Na účely tohto projektu potrebujeme mať značnú históriu volebných prieskumov v jednotnom formáte. Nakoľko však agentúra FOCUS nezverejňuje s každým prieskumom jednotný .xlsx alebo .csv súbor tak je táto úloha obzvlášť problematická. S každým vykonaným prieskumom majú k dispozícii .pdf v štýle reportu (Press release), pri ktorých sme skontrolovali viaceré rôzne a javí sa, že FOCUS pri týchto reportoch udržuje jednotný formát v priebehu rokov. Súčasťou každého takéhoto .pdf súboru je aj samotný prieskum, napríklad:

Politická strana	**% rozhodnutých voličov**	**95% interval spoľahlivosti**	% rozhodnutýchrozhodnutých voličov	% voličov
	september 2019		august 2019 II.	august 2019 I.
SMER-SD	**21,7%**	18,8% - 24,6%	20,8%	21,8%
koalícia strán Progresívne Slovensko a SPOLU	**13,3%**	10,9% - 15,7%	14,9%	14,0%
Kotleba – Ľudová strana Naše Slovensko	**10,6%**	8,4% - 12,8%	11,4%	12,1%
SME RODINA – Boris Kollár	**7,2%**	5,4% - 9,0%	6,6%	6,3%
KDH	**6,9%**	5,1% - 8,7%	7,3%	7,5%
SNS	**6,8%**	5,0% - 8,6%	7,0%	7,0%
OĽANO	**6,8%**	5,0% - 8,6%	6,1%	6,0%
Za ľudí	**6,5%**	4,7% - 8,3%	5,5%	5,0%
SaS	**6,4%**	4,6% - 8,2%	7,9%	7,0%
MOST-HÍD	**4,1%**	2,7% - 5,5%	4,3%	4,7%
SMK – MKP	**3,3%**	2,1% - 4,5%	3,5%	3,4%
Dobrá voľba	**2,1%**	-	x	x
SZS	**1,0%**	-	0,3%	0,9%
iná strana[3]	**3,3%**	-	4,4%	4,3%

** v tabuľke sú uvedené len subjekty, ktoré dosiahli preferencie 1% a viac*

Pustili sme sa teda do automatizovaného sťahovania zo stránok FOCUSu pomocou python knižnice Selenium. Podarilo sa nám takýchto reportov získať 127.

2.2 Extrahovanie údajov

Ďalej sme hľadali spôsob ako tieto tabuľky automatizovane extrahovať na účel vytvorenia jednotného .csv súboru. Použili sme Python knižnicu Docling, pomocou ktorej sme prekonvertovali tabuľky z .pdf do pd.DataFrame, spojili a exportovali do jednotného .csv.

Tento proces nebol úplne priamočiary. Názvy politických strán sa extrahovali veľmi nekonzistentne. Nie len to ale aj fakt, že viaceré politické strany menili meno v priebehu rokov. Napríklad strana Progresívne Slovensko v roku 2020 kandidovala ako koalícia so stranou SPOLU, avšak v ďalších voľbách už kandidovala iba ako Progresívne Slovensko. Podobných scénarov bolo viacero, takže sme manuálne vytvorili maper, ktorý tieto názvy zjednotil. Ďalej bolo potreba niektoré záznamy aj manuálne opraviť, keďže strany s dlhším názvoch (v reporte zabrali dva riadky tabuľky) sa duplikovali v našom extrahovanom datasete.

Po týchto všetkých úkonoch sme už mali jednotný formát dát, ktorý zachytával prieskumy z mnohých mesiacov.

3 Exploratívna analýza

4 Klasifikačné modely

5 Predikčné modely

```
class MyClass(Yourclass):
  def __init__(self, my, yours):
    bla = '5 1 2 3 4'
    print bla
```

6 Predikcie volieb

7 Zhrnutie