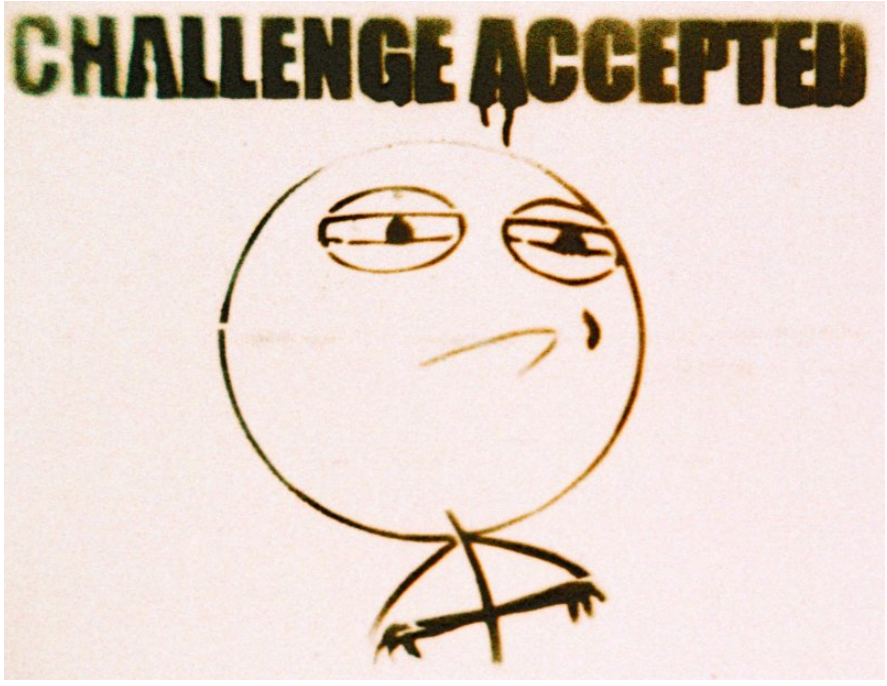
Duomenų analizė naudojant Python

Mantas Zimnickas

PyCon-LT 2013

Duomenų analizės užduotis:

Palyginti dvi populiarias seimo frakcijas tarpusavyje.



http://www.flickr.com/photos/schoeband/8205785361/

Atsakymas į gyvybę, visatą ir viską:

42

Atsakymas į dviejų populiarių seimo frakcijų tarpusavio palyginimą:

-0.24141722277698799

LIETUVOS RESPUBLIKOS SEIMAS

WWW.LRS.LT

Pasirengimas Lietuvos valstybės atkūrimo šimtmečiui

Lietuvos Respublikos Seimas skelbia Kalbos premijos konkursą

SEIMO POSĖDŽIAI

Paskutinis įvykęs Seimo posėdis
Informacija apie Seimo posėdžius
Seimo sesijos

PARLAMENTINĖ KONTROLĖ

RENGINIAI

PRANEŠIMAI ŽINIASKLAIDAI

Užsienio reikalų komitetas apsvarstė Seimo Statuto pakeitimo projektą 2013-04-26 15:13
Oficiali Seimo Pirmininko Vydo Gedvilo 2013 m. balandžio 29 d., pirmadienio, darbotvarkė 2013-04-26 14:14
Mirė VI Seimo narys Romualdas Ignas BLOŠKYS 2013-04-26 13:16

TEISĖKŪRA

- Teisės aktų ir Seime įregistruotų teisės aktų projektų paieška
- Valstybės ir savivaldybių institucijų teisės aktų projektų paieška
- ES teisės aktų paieška
- 1918-1940 m. teisės aktų paieška
- Konstitucija
- Seimo statutas
- Kodeksai
- Naujienos

EN

Seimo sesijos

Pavadinimas	Pradžia	Pabaiga
2012 - 2016 metų kadencija		
2 eilinė sesija	2013-03-10	
1 eilinė sesija	2012-11-16	2013-01-17
2008 - 2012 metų kadencija		
9 eilinė sesija	2012-09-10	2012-11-14
9 neeilinė sesija	2012-07-16	2012-07-16
8 eilinė sesija	2012-03-10	2012-06-30
8 neeilinė sesija	2012-01-30	2012-01-30
7 neeilinė sesija	2012-01-17	2012-01-19
7 eilinė sesija	2011-09-10	2011-12-23
6 eilinė sesija	2011-03-10	2011-06-30
5 eilinė sesija	2010-09-10	2010-12-23
4 eilinė sesija	2010-03-10	2010-07-02
3 neeilinė sesija	2010-02-11	2010-02-11
3 eilinė sesija	2009-09-10	2010-01-21
2 eilinė sesija	2009-03-10	2009-07-23
2 neeilinė sesija	2009-02-05	2009-02-19

EN

Sesijos pasirinkimas

8 eilinė Seimo sesija (2012-03-10 - 2012-06-30)

Diena	Posėdžiai
2012-06-30	rytinis
2012-06-29	rytinis neeilinis, vakarinis neeilinis
2012-06-28	rytinis, vakarinis
2012-06-26	rytinis, vakarinis
2012-06-21	rytinis, vakarinis
2012-06-20	neeilinis
2012-06-19	rytinis, vakarinis
2012-06-14	vakarinis
2012-06-12	rytinis, vakarinis
2012-06-07	rytinis, vakarinis
2012-06-05	rytinis, vakarinis
2012-05-24	rytinis, vakarinis
2012-05-22	rytinis, vakarinis
2012-05-17	rytinis, vakarinis
2012-05-15	rytinis, vakarinis
2012-05-10	rytinis, vakarinis

EN

8 eilinė Seimo sesija

Seimo posėdis Nr.460 (2012-06-28, vakarinis)

- Protokolas
- Stenograma
- Garso įrašas
- Lankomumas

Laikas	Numeris	Svarstytas klausimas
15:02	1 - 13.	Pridėtinės vertės mokesčio įstatymo 19 straipsnio papildymo ir pakeitimo ĮSTATYMO PROJEKTAS (Nr. XIP-3885GR) [Grąžinto įstatymo pateikimas]
15:26	1 - 14.	Mėgėjiškos žūklės įstatymo pakeitimo ĮSTATYMO PROJEKTAS (nauja redakcija) (Nr. XIP-45GR) [Grąžinto įstatymo pateikimas]
16:00	1 - 16a.	Lietuvos kultūros tarybos ĮSTATYMO PROJEKTAS (Nr. XIP-3469(4)) [Svarstymas]
16:03	2 - 1.	Etikos ir procedūrų komisijos išvada Dėl Seimo laikinosios tyrimo komisijos darbo
16:18	1 - 11.	Šilumos ūkio įstatymo 20 straipsnio pakeitimo ĮSTATYMO PROJEKTAS (Nr. XIP-4210(5)) [Svarstymas]
16:34	1 - 11.	Šilumos ūkio įstatymo 20 straipsnio pakeitimo ĮSTATYMO PROJEKTAS (Nr. XIP-4210(5)) [Priėmimas]
16:44	2 - 2e.	Pensijų sistemos reformos įstatymo 1, 2, 3, 4, 7 ir 8 straipsnių pakeitimo ĮSTATYMO PROJEKTAS (Nr. XIP-3381(3)) [Svarstymas]
16:46	2 - 2.	Klausimų grupė: 2 - 2a, 2 - 2b, 2 - 2c, 2 - 2d, 2 - 2e, 2 - 2f [Svarstymas]
17:33	2 - 5.	Įstatymo "Dėl užsieniečių teisinės padėties" 1, 2, 6, 9, 10, 11, 12(1), 17, 19, 21, 22, 24, 26, 33, 37, 38, 40, 43, 49(1), 50, 53, 54, 55, 57, 58, 89, 97, 98, 99, 100, 101, 102, 104, 106, 113, 128, 131, 133, 139, 140(1), 141(1) straipsnių ir priedo pakeitimo ir papildymo, Įstatymo papildymo 44(1), 49(3), 98(1), 99(1), 103(1), 105, 105(1), 105(2), 105(3), 105(4), 106(1) straipsniais ir 12(2), 13, 14, 15, 16, 18, 20, 145 straipsnių pripažinimo netekusiais galios ĮSTATYMO PROJEKTAS (Nr. XIP-2360(3)) [Svarstymas]

LIETUVOS RESPUBLIKOS SEIMAS

EN

8 eilinė Seimo sesija

Seimo posėdis Nr. 460 (2012-06-28, vakarinis)

Darbotvarkės klausimas

Šilumos ūkio įstatymo 20 straipsnio pakeitimo įSTATYMO PROJEKTAS (Nr. XIP-4210(5)); priėmimas
(dokumento tekstas, susiję dokumentai)
Pranešėjai:
Dainius Budrys, Komiteto pirmininkas, Ekonomikos komitetas, Lietuvos Respublikos Seimas,
Danutė Bekintienė, Komiteto narė, Valstybės valdymo ir savivaldybių komitetas, Lietuvos Respublikos Seimas

Svarstymo eiga

16:36:06	Kalbėjo Asta Baukutė
16:36:53	Kalbėjo Edmundas Pupinis
16:38:20	Kalbėjo Egidijus Klumbys
16:40:32	Kalbėjo Edvardas Žakaris
16:42:03	Kalbėjo Kęstutis Masiulis
16:42:38	Kalbėjo Vaidotas Bacevičius
16:43:06	Įvyko registracija (užsiregistravo **76**)
16:43:06	Įvyko balsavimas dėl įstatymo priėmimo; **pritarta** (už **75**, prieš **0**, susilaikė **1**)

EN

8 eilinė Seimo sesija

Seimo posėdis Nr. 460 (2012-06-28, vakarinis)

Balsavimo rezultatai

Darbotvarkės klausimas

Šilumos ūkio įstatymo 20 straipsnio pakeitimo ĮSTATYMO PROJEKTAS (Nr. XIP-4210(5)); priėmimas
(dokumento tekstas, susiję dokumentai)
Pranešėjai:
Dainius Budrys, Komiteto pirmininkas, Ekonomikos komitetas, Lietuvos Respublikos Seimas,
Danutė Bekintienė, Komiteto narė, Valstybės valdymo ir savivaldybių komitetas, Lietuvos Respublikos Seimas

Formuluotė: **dėl įstatymo priėmimo**
Balsavimo laikas: **16:43:06**
Balsavo Seimo narių: **76** iš **139**.
Balsavimo rezultatai: už - **75**, prieš - **0**, susilaikė - **1**, **pritarta**.

Pateikti balsavimo rezultatus pagal frakcijas

Individualūs balsavimo rezultatai

Seimo narys(-ė)	Frakcija	Už	Prieš	Susil.
Adomėnas Mantas	TSLKDF	+		
Aleknaitė Abramikienė Vilija	TSLKDF	+		
Andriukaitis Vytenis Povilas	LSDPF	+		
Anušauskas Arvydas	TSLKDF	+		
Auštrevičius Petras	LSF	+		
Ažubalis Audronius	TSLKDF	+		

Kaip gauti duomenis?

Atrodo greičiausias būdas - **paprašyti**.

- Skambinau seimo kanceliarijos informacijos technologijų ir telekomunikacijų departamento direktoriaus pavaduotojui.
- Parašiau oficialų pareiškimą dėl galimybės gauti duomenis.
- Vis dar laukiu atsakymo.

Kaip gauti duomenis?

Scrapy - bandžiau, atsisakiau.

- Prastas puslapių kešavimas.
- Nėra puslapių atsiuntimo eilės valdymo.
- Kreivas `Item` objektų įgyvendinimas.
- Sunku debuginti.
- Ir dar, sunku debuginti.

Kaip gauti duomenis?

databot - mano bandymas padaryti geresnį Scrapy.

https://bitbucket.org/sirex/databot

Kaip veikia databot

1. Rašomos atsiuntimo instrukcijos.
2. Atsiunčiami ir išsaugomi puslapiai.
3. Rašomas/taisomas parseris.
4. Leidžiamas parseris visiems puslapiams
5. Kartojami 3 ir 4 žingsniai, kol parseris pradeda veikti su visais puslapiais.

Saugojimo problema

```
pages/www3.lrs.lt/
pages/www3.lrs.lt/79e/
pages/www3.lrs.lt/79e/6c7/
pages/www3.lrs.lt/79e/6c7/f76/
pages/www3.lrs.lt/79e/6c7/f76/a1581e15dcd3b711577d406
pages/www3.lrs.lt/79e/722/
pages/www3.lrs.lt/79e/722/a8a/
pages/www3.lrs.lt/79e/722/a8a/1635e66270a093483ff683a
pages/www3.lrs.lt/79e/ba9/
pages/www3.lrs.lt/79e/ba9/efe/
pages/www3.lrs.lt/79e/ba9/efe/45934172d21653aa83250dc
pages/www3.lrs.lt/79e/74f/
pages/www3.lrs.lt/79e/74f/5bf/
pages/www3.lrs.lt/79e/74f/5bf/391d3ba0ac14bace5817db1
pages/www3.lrs.lt/79e/178/
```

Saugojimo problema

Testas su ~32 000 puslapių.

```
total = 0
for page in Page.objects.all():
    path = page.get_file_path()
    with open(path) as f:
        content = f.read()
    total += len(content)
print total

# $ time ./hashfiles.py
# 1558376577
# ./hashfiles.py  6,45s user 6,06s system 3% cpu 6:35,38 total
```

Saugojimo problema

Testas su ~32 000 puslapių.

```
import msgpack
total = 0
with open('out') as f:
    for x, x, content in msgpack.Unpacker(f):
        total += len(content)
print total

# $ time ./msgpack.py
# 1558376577
# ./msgpack.py  1,55s user 1,25s system 16% cpu 16,619 total
```

Koduočių pokštas

Visi lrs.lt puslapiai:

```
<meta content="text/html; charset=windows-1257"
<meta content="Lietuvos Respublikos Seimo oficia
Internete" name="Description">
```

Tikroji koduotė:

```
parser = lxml.html.HTMLParser(encoding='iso-8859-13')
html = lxml.html.parse(content, parser)
```

```
$ du -sh votes.csv
75M     votes.csv

$ head -n3 votes.csv
bals_id,fraction,asm_id,vote
-15445,DKF,50,0
-15445,DKF,7196,0

$ wc -l votes.csv
 4309030 votes.csv
```

Skaičiai

Balsas	Skaitinė reikšmė
Už	2
Susilaikė	-1
Prieš	-2

```
[pylab]
recipe = zc.recipe.egg
eggs =
    numpy
    scipy
    matplotlib
    numexpr
    tables
    pandas
    ipython
    networkx
    nltk
    MDP
```

```
$ ipython --pylab
Welcome to pylab, a matplotlib-based Python
environment [backend: GTKAgg].
For more information, type 'help(pylab)'.

$ ipython notebook
The IPython Notebook is running at:

    http://127.0.0.1:8888/
```

```
>>> import numpy as np
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
```

```
>>> d = pd.read_csv('votes.csv')

>>> d.head()
   bals_id fraction  asm_id  vote
0   -15445      DKF      50     0
1   -15445      DKF    7196     0
2   -15445      DKF   53917     2
3   -15445      DKF   56356     2
4   -15445      DKF   73590     2
```

Atminties naudojimas:

```
>>> d.dtypes
bals_id      int64
fraction    object
asm_id       int64
vote         int64

>>> d.shape, d.ix[0].nbytes
((4309029, 4), 32)

>>> d.shape[0] * d.ix[0].nbytes / 1024**2
131
```

```
$ ps o rss,args | grep pylab
63616 ipython --pylab

$ ps o rss,args | grep pylab
232772 ipython --pylab

$ echo $((232772-63616))
169156
```

rss

resident set size, the non-swapped physical memory that a task has used (in kiloBytes).

numpy

```
>>> a = np.random.randn(6, 2) ; a
array([[ 0.11, -0.74],
       [-0.97, -0.56],
       [ 1.07,  1.42],
       [-0.4 ,  2.43],
       [-0.4 ,  0.05],
       [-0.01,  1.66]])
```

numpy

```
>>> a + a
array([[ 0.22, -1.49],
       [-1.94, -1.12],
       [ 2.14,  2.84],
       [-0.79,  4.86],
       [-0.8 ,  0.1 ],
       [-0.03,  3.33]])
```

numpy

```
>>> np.abs(a)
array([[ 0.11,  0.74],
       [ 0.97,  0.56],
       [ 1.07,  1.42],
       [ 0.4 ,  2.43],
       [ 0.4 ,  0.05],
       [ 0.01,  1.66]])
```

numpy

```
>>> a[0]
array([ 0.11, -0.74])

>>> a[:,0]
array([ 0.11, -0.97,  1.07, -0.4 , -0.4 , -0.01])

>>> a[a>0]
array([ 0.11,  1.07,  1.42,  2.43,  0.05,  1.66])
```

numpy

```
>>> a > 0
array([[ True, False],
       [False, False],
       [ True,  True],
       [False,  True],
       [False,  True],
       [False,  True]], dtype=bool)
```

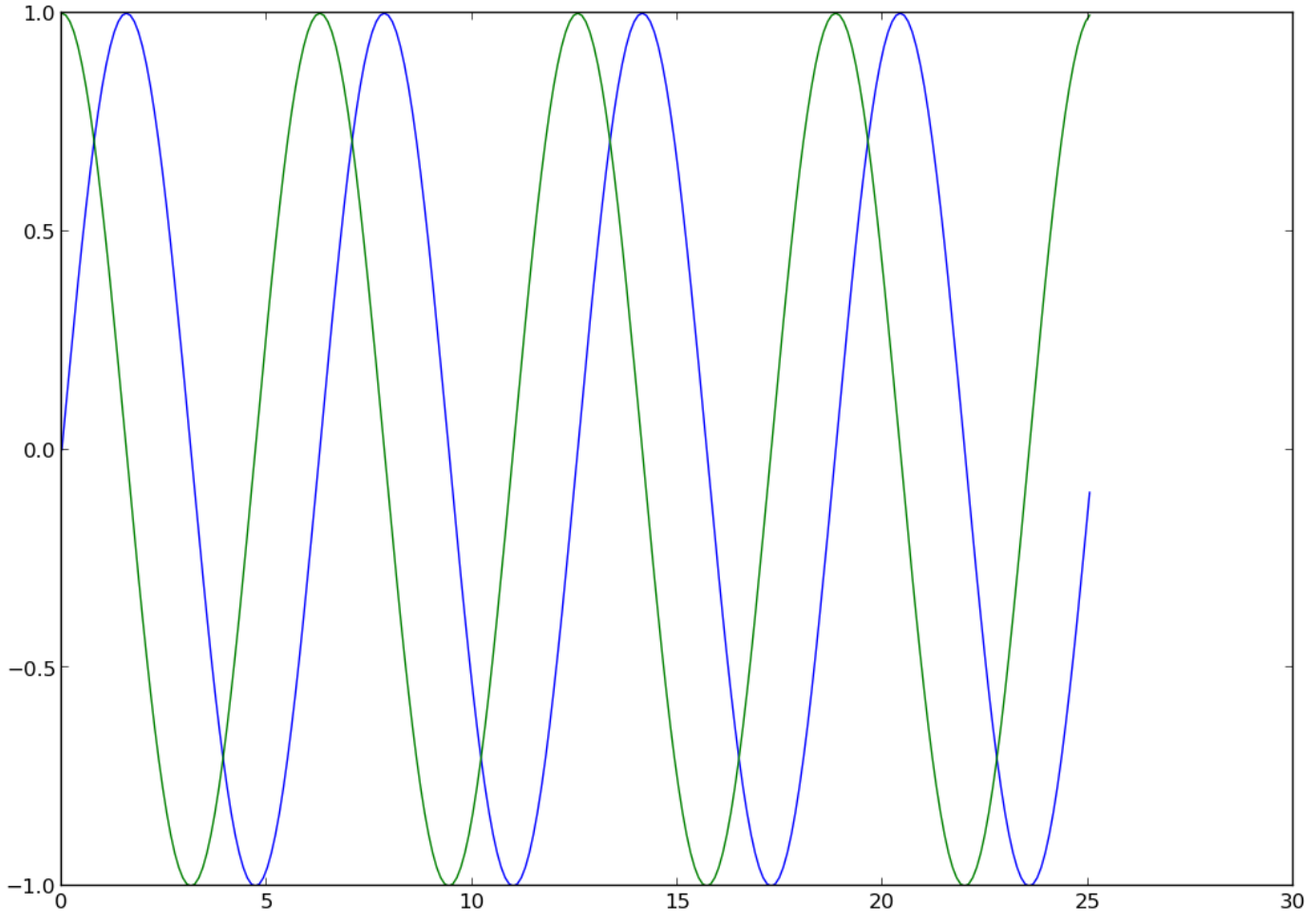
numpy

```
>>> a.dtype
dtype('float64')

>>> a.nbytes
96

>>> (a>0).nbytes
12
```

numpy

```
>>> a.reshape((4, 3))
array([[ 0.11, -0.74, -0.97],
       [-0.56,  1.07,  1.42],
       [-0.4 ,  2.43, -0.4 ],
       [ 0.05, -0.01,  1.66]])
```

```
>>> x = np.arange(0, 8*np.pi, np.pi/32)

>>> y = np.sin(x)

>>> plt.plot(x, y)
[<matplotlib.lines.Line2D at 0xb870910>]

>>> y = np.cos(x)

>>> plt.plot(x, y)
[<matplotlib.lines.Line2D at 0x9992250>]
```


1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
30

```
>>> idx = ['bals_id', 'fraction', 'asm_id']

>>> d = d.set_index(idx)

>>> d = pd.read_csv('votes.csv', index_col=idx)

>>> d.head()
                          vote
bals_id fraction asm_id
-15445  DKF      50          0
                 7196        0
                 53917       2
                 56356       2
                 73590       2

>>> d = d.sort()
```

```
>>> d.ix[(-15445, 'DKF')]
         vote
asm_id
50          0
7196        0
53917       2
56356       2
73590       2
73591       0
73592       0

>>> d.ix[(-15445, 'DKF')].mean()
0.857143
```

```
>>> d = d.mean(level=('bals_id', 'fraction'))

>>> d.head()
                     vote
bals_id fraction
-15445  DKF       0.857143
        DPF       1.310345
        LLRAF     2.000000
        LSDPF     1.157895
        LSF       0.400000

>>> d.shape
(258391, 1)
```

```
>>> d = d.unstack()

>>> d.shape
(31096, 34)

>>> d[d.columns[:4]].head()
                vote
fraction         AŽF  CF        DKF        DPF
bals_id
-15445           NaN NaN  0.857143  1.310345
-15444           NaN NaN  0.857143  1.379310
-15443           NaN NaN  0.857143  1.241379
-15442           NaN NaN  0.857143  1.103448
-15441           NaN NaN  0.857143  1.448276
```

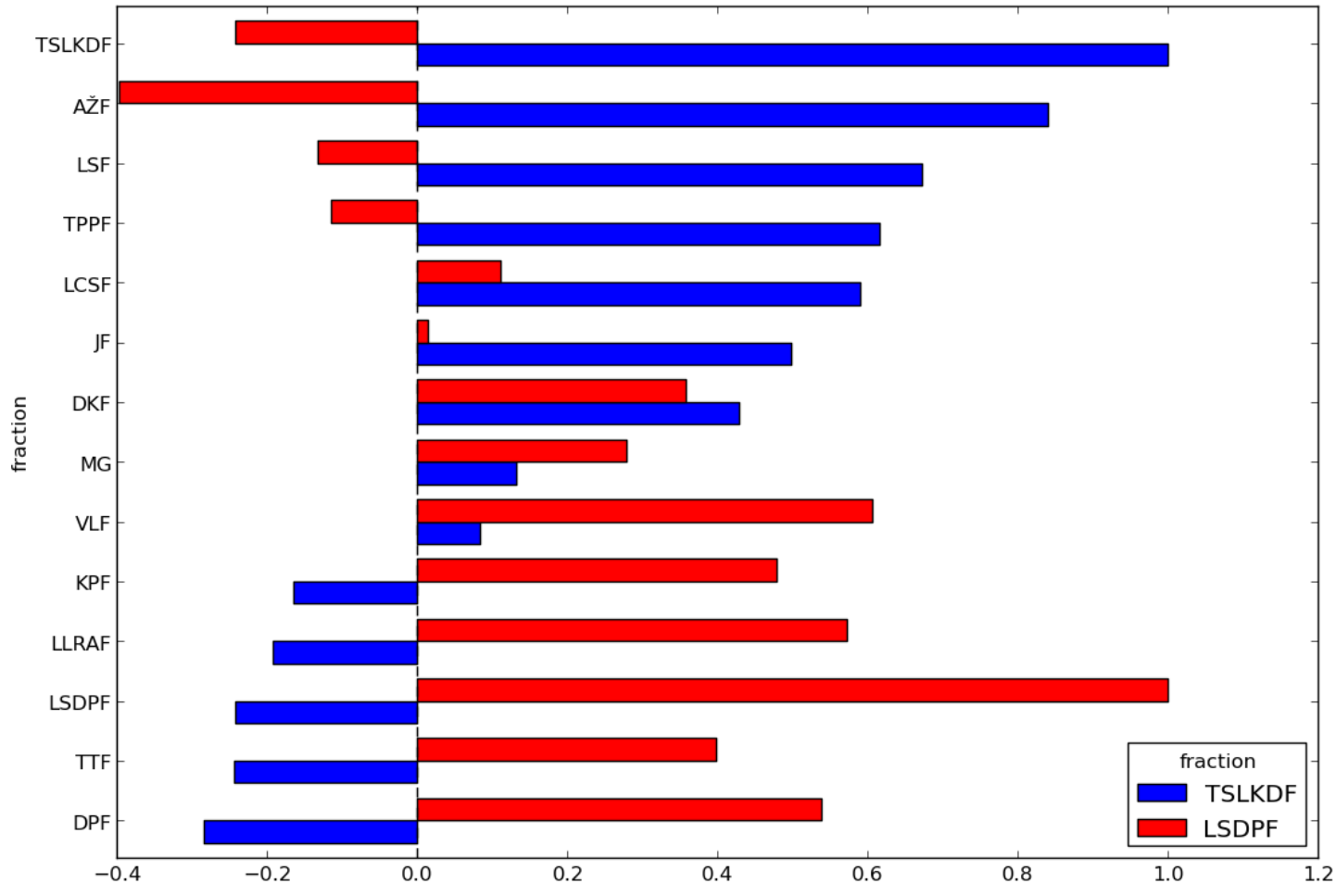
```
>>> d.columns.get_level_values('fraction')
Index([AŽF, CF, DKF, DPF, JDTLF, JF, JLF, KDF,
       KPF, LCSF, LDDP, LDF, LF, LLRAF, LSDPF,
       LSF, MG, MKDF, NF, NKF, NSF, PDF, SDF,
       SDF2000, SDKF, TPPF, TSF, TSKF, TSLK,
       TSLKDF, TTF, VLF, VLPD, VNDF],
       dtype=object)

>>> d.columns = d.columns.get_level_values('fraction')
```

```
>>> d[d.columns[:4]].head()
fraction  AŽF  CF        DKF        DPF
bals_id
-15445    NaN NaN  0.857143  1.310345
-15444    NaN NaN  0.857143  1.379310
-15443    NaN NaN  0.857143  1.241379
-15442    NaN NaN  0.857143  1.103448
-15441    NaN NaN  0.857143  1.448276
```

```
>>> d.count().order(ascending=False).head(10)
fraction
MG          29986
LSDPF       18088
DPF         18088
LCSF        18024
NSF         16893
LSF         15973
TTF         15734
JF          12029
VNDF        10804
TSLKDF      10424
```

```
>>> d.corr()[['TSLKDF', 'LSDPF']].dropna() \
...   .sort_index(by='TSLKDF')
fraction    TSLKDF     LSDPF
fraction
DPF       -0.283767  0.539405
TTF       -0.243230  0.398492
LSDPF     -0.241417  1.000000
LLRAF     -0.191360  0.572678
KPF       -0.165205  0.479061
VLF        0.084323  0.605972
MG         0.132368  0.279152
DKF        0.428781  0.357727
JF         0.498740  0.014144
LCSF       0.589915  0.111371
TPPF       0.615763 -0.114505
```

```
>>> _.plot(kind='barh')
```


TSLKDF
AŽF
LSF
TPPF
LCSF
JF
DKF
MG
VLF
KPF
LLRAF
LSDPF
TTF
DPF
fraction
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
fraction
TSLKDF
LSDPF

```
TSLKDF  Tėvynės sąjungos-Lietuvos krikščionių demokratų frakcija
AŽF     "Ąžuolo" frakcija
LSF     Liberalų sąjūdžio frakcija
TPPF    Tautos prisikėlimo partijos frakcija
LCSF    Liberalų ir centro sąjungos frakcija
JF      Jungtinė frakcija
DKF     Frakcija "Drąsos kelias"
MG      Mišri Seimo narių grupė
VLF     Frakcija "Viena Lietuva"
KPF     Krikščionių partijos frakcija
LLRAF   Lietuvos lenkų rinkimų akcijos frakcija
LSDPF   Lietuvos socialdemokratų partijos frakcija
TTF     Frakcija "Tvarka ir teisingumas"
DPF     Darbo partijos frakcija
```

```
>>> d = d[['LSDPF','TSLKDF']]

>>> d.head()
fraction      LSDPF    TSLKDF
bals_id
-15445    1.157895  0.939394
-15444    1.105263  0.848485
-15443    1.052632  0.969697
-15442    1.052632  1.030303
-15441    1.157895  1.060606
```

```
>>> d.describe()
fraction          LSDPF         TSLKDF
count      18088.000000  10424.000000
mean           0.417052      0.760099
std            0.806694      0.981152
min           -2.000000     -2.000000
25%           -0.190476      0.021739
50%            0.565217      1.155556
75%            1.052632      1.511111
max            2.000000      2.000000
```

```
>>> (d.LSDPF - d.TSLKDF).std()
1.4389501063185526
```

```
>>> d = d.dropna()

>>> d.count()
fraction
LSDPF        10424
TSLKDF       10424
```

```
>>> d.corr()
fraction      LSDPF    TSLKDF
fraction
LSDPF      1.000000 -0.241417
TSLKDF    -0.241417  1.000000

>>> d.LSDPF.corr(d.TSLKDF)
-0.24141722277698799
```

Ačiū už dėmesį.