Python data alkalmazások a CEU-n

Zágoni Rita
programozó, CEU MicroData

zagonir@ceu.hu
twitter.com/ritazagoni

Rólunk

- CEU MicroData: közgazdaságtani kutatócsoport
- kutatók, programozók, elemzők
- szöveg,-és adatfeldolgozás, elemzés
- alkalmazások, adatok tanulmányokhoz

kozbeszerzes.ceu.hu

Kutatási témák

- cégháló
- cégek politikai kapcsolatai
- menedzsment-adatok

Feladatok

- adat kinyerése – pl. webről
- adatbázisok formába hozása, szabad szöveges mezőkből strukturált adatok
- adatbázisok összekötése
- általános **Python**
- ökonometriai elemzés: **Stata**
- tisztítás, predikció**: pandas, scikit-learn**

Adatkezelés

- Pandas
- adattisztítás
- formázás
- leíró statisztika

Adattisztítás

- hiányzó értékek
- cleaned = df.dropna() - NaN-t tartalmazó sorok/oszlopok törlése
- default: minden sort, amely tartalmaz hiányzó értéket
- vagy csak ahol minden érték NA

Adattisztítás

- fillna() - konstans, kitöltés előre
- df.fillna(‘unknown’)
- df.fillna(df.mean())
- üres stringek behelyettesítése:
- df[df.name == ''] = 'unknown'

Adattisztítás

- duplikátumok törlése
- df.drop duplicates() megadható, hogy melyik oszlop(ok) alapján

Adatelemzés

- leíró statisztika
- mindegyik statisztikai függvény figyelmen kívül hagyja a hiányzó adatokat
- df.position.value_counts()

- hr 1328
- team leader 762
- general manager 312
- director 216
- CEO 20

Adatelemzés

- df.describe()
- darabszám, átlag, szórás, min, max, kvantilisek

count	500.000000
mean	0.481243
std	0.288896
min	0.000257
25%	0.229442
50%	0.462463
75%	0.734411
max	0.997569

Hiányzó értékek/predikció

nemzetiségi kitalálása név alapján: scikit-learn

- probléma
  - menedzserek nemzetiségi hovatartozása – hiányosan kitöltött
- megoldás:
  - a kitöltöttek mint tréningadat – nemzetiség kitalálása új nevek esetén

szöveges input:

name	nationality
Peter Smith	US
Walter Müller	DE
Jacques Leloup	FR
Luca Antonioli	IT

```
import pandas as pd
Import numpy as np

df = pd.read_csv(“name_nationality.csv”,
encoding='UTF-8')
df = df[df.notnull()]
data =
df.reindex(np.random.permutation(df.index))
data = data.groupby(data['nationality'],
as_index=False).filter(lambda x: len(x['name']) >
100)
```

- beolvassa a tréningadatokat tartalmazó csv-t
- adatkezelés: pandas dataframe

```
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.linear_model import SGDClassifier
from sklearn.pipeline import Pipeline
from sklearn.cross_validation import
train_test_split

pipeline = Pipeline([
    ('vectorizer',   CountVectorizer(ngram_range=(1,
4),                              analyzer='char')),
('classifier',  SGDClassifier())])
x_train, x_test, y_train, y_test =
train_test_split(data['name'], data['nationality'],
test_size=0.2, random_state=1)
```

- kinyeri a karakterisztikus jegyeket: karakter n-gramok
- betanít egy klasszifikáló algoritmust
- tréning,- és tesztadat
- új nevekre megpróbálja kitalálni a nemzetiségüket
- nem optimalizált

Miért Python

- szerteágazó feladatokra
- általános célú programnyelv
- adatelemzési eszköztárak
- nem kell
  - nyelvek
  - környezetek, formátumok
  - között váltani

blog.defacto.io

blog.microdata.io

CEU MICRODATA

SEPTEMBER 11, 2014

Reproducible plumbing

by *Miklós Koren*

> "I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis." (interviewee in the seminal Kandel, Paepcke, Hellerstein and Heer interview study of business analytics practices)

In fact, my estimate is that about 80 percent of the work I do in an empirical research project is about getting, transforming, merging, or otherwise preparing data for the actual analysis.

Kérdések?

Köszönöm a figyelmet!