Statystyka użytkowników serwisu bit.ly

Projekt z Technik Komputerowych w Fizyce

Michał Kaczmarczyk, Maciej Dudek, Alexandros Chantelidis

1 Wstęp

Celem projektu jest przedstawienie statystki użytkowników serwisu skracającego linki bit.ly w zależności od strefy czasowej, w jakiej się znajdują, jak i przeglądarki, której używają. Dane są aktualizowane na stronie usa.gov co godzinę.

2 Kod programu

Program jest napisany w całości w iPythonie w środowisku Jupyter, z użyciem bibliotek numpy, matplotlib i pandas.
Kolejność kroków w programie jest następująca:

1. **Import potrzebnych bibliotek**

```
In [3]: #import bibliotek
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import json
from pandas import DataFrame, Series
```

2. **Wczytanie pliku i wyświetlenie jego zawartości**

```
In [4]: #wczytanie pliku
path = 'usagov_bitly_data2012-03-16-1331923249.txt'
records = [json.loads(line) for line in open(path)]
```

```
In [5]: #wyświetlenie zawartości pliku
records[0]
```

```
Out[5]: {'a': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.11
(KHTML, like Gecko) Chrome/17.0.963.78 Safari/535.11',
'al': 'en-US,en;q=0.8',
'c': 'US',
'cy': 'Danvers',
'g': 'A6qOVH',
'gr': 'MA',
'h': 'wfLQtf',
'hc': 1331822918,
'hh': '1.usa.gov',
'l': 'orofrog',
'll': [42.576698, -70.954903],
'nk': 1,
'r': 'http://www.facebook.com/l/7AQEFzjSi/1.usa.gov/wfLQtf',
't': 1331923247,
'tz': 'America/New_York',
'u': 'http://www.ncbi.nlm.nih.gov/pubmed/22415991'}
```

Plik ma kilkanaście wpisów odnośnie użytkowników serwisu bit.ly, jak ich język, lokalizację, strefę czasową, system operacyjny, jaki używają itd.

3. **Wybranie kolumny ze strefami czasowymi i ich zliczenie**

```
In [6]: #pierwsze 10 stref czasowych
frame = DataFrame(records)
frame['tz'][:10]
```

```
Out[6]: 0     America/New_York
1       America/Denver
2     America/New_York
3    America/Sao_Paulo
4     America/New_York
5     America/New_York
6        Europe/Warsaw
7
8
9
Name: tz, dtype: object
```

```
In [7]: #zliczenie stref czasowych i wyświetlenie najbardziej popularnych
tz_counts = frame['tz'].value_counts()
tz_counts[:15]
```

```
Out[7]: America/New_York        1251
521
America/Chicago          400
America/Los_Angeles      382
America/Denver           191
Europe/London             74
Asia/Tokyo                37
Pacific/Honolulu          36
Europe/Madrid             35
America/Sao_Paulo         33
Europe/Berlin             28
Europe/Rome               27
America/Rainy_River       25
Europe/Amsterdam          22
America/Indianapolis      20
Name: tz, dtype: int64
```

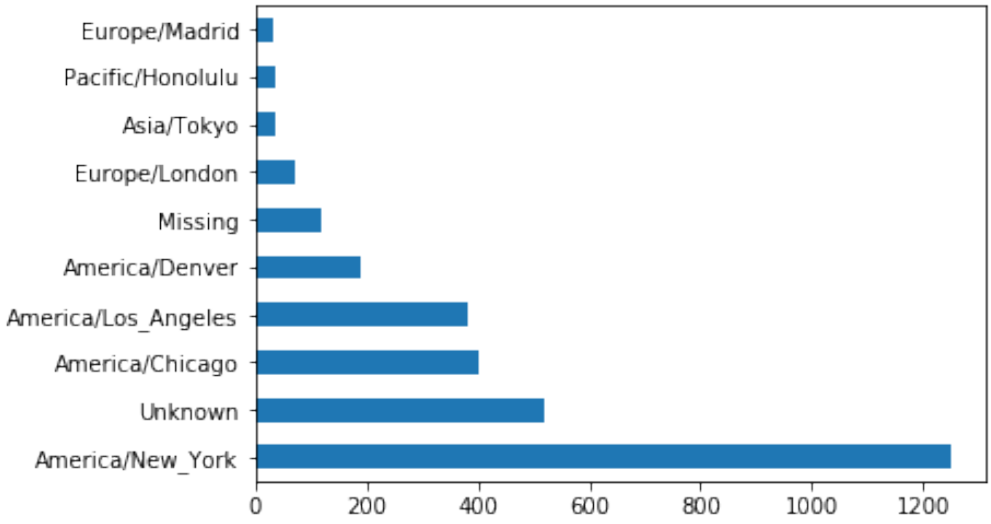
4. **Zastąpienie nieistniejących danych frazą "Unknown"**

```
In [8]: #czyszczenie nieistnejących wartości
clean_tz = frame['tz'].fillna('Missing')
clean_tz[clean_tz == ''] = 'Unknown'
tz_counts = clean_tz.value_counts()
tz_counts[:3]
```

```
Out[8]: America/New_York    1251
Unknown               521
America/Chicago       400
Name: tz, dtype: int64
```

5. **Sporządzenie wykresu najbardziej popularnych stref czasowych**

```
In [9]: #Wykres najbardziej popularnych stref czasowych
tz_counts[:10].plot(kind='barh')

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x74ca2f0>
```

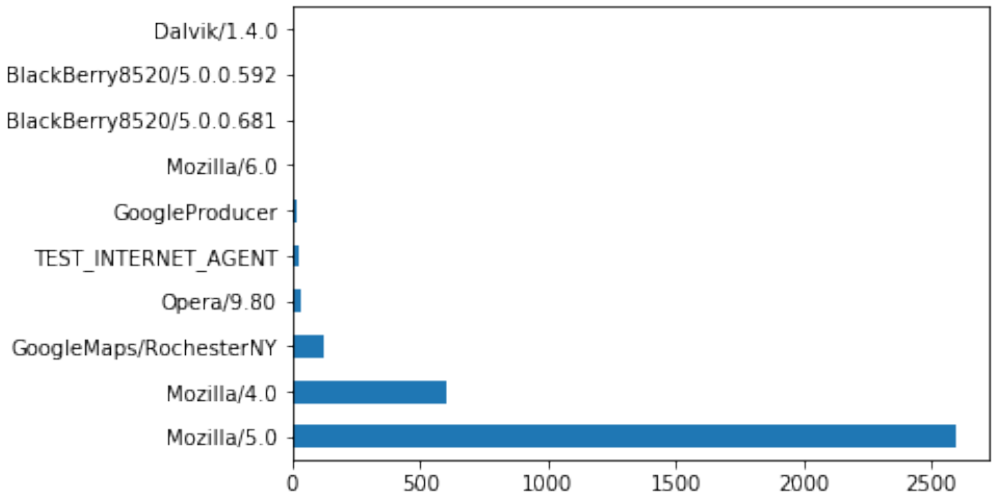


6. **Podzialenie kolumny ”a” na części (zawiera dane o przeglądarce)**

```
In [10]: #Statystyka przeglądarek użytkowników
results = Series([x.split()[0] for x in frame.a.dropna()])
results[:5]

Out[10]: 0               Mozilla/5.0
1    GoogleMaps/RochesterNY
2               Mozilla/4.0
3               Mozilla/5.0
4               Mozilla/5.0
dtype: object

In [11]: #Top 10 przeglądarek
results.value_counts()[:10]

Out[11]: Mozilla/5.0                 2594
Mozilla/4.0                  601
GoogleMaps/RochesterNY       121
Opera/9.80                    34
TEST_INTERNET_AGENT           24
GoogleProducer                21
Mozilla/6.0                    5
BlackBerry8520/5.0.0.681       4
BlackBerry8520/5.0.0.592       3
Dalvik/1.4.0                   3
dtype: int64
```

7. **Sporządzenie wykresu używanych przeglądarek**

```
In [12]: #Wykres przeglądarek
results.value_counts()[:10].plot(kind='barh', rot=0)

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0xa1f2d90>
```



3 Przyszłość programu

Program ma możliwość robienia statystyki ze wszystkich zawartych danych w pliku. Można go rozwijać przez dodanie możliwości sporządzenia statystyki używanego systemu operacyjnego, czy też można nanieść lokalizacje użytkowników na mapę.