Scraping Angel.co

All Companies

DONE DEALS RETURNS BROWSE ALL

All Companies 4497K

List ▾ Type ▾ Location ▾ Market ▾ Tech ▾ Investors ▾ Team ▾ Stage ▾ Ranges ▾

4,497,128 Companies

Company	Signal	Joined	Location	Market	Website	Employees	Stage	Total Raised
Prayas Analytics A/B testing for stores.		Apr '14	New York City	Retail Technology	prayasanalytic…	1-10	-	-
WOO Sports On-board motion sensor that turns…		Apr '14	Boston	Gamification	woosports.com	1-10	Series A	$4,300,000

Angel-Hack.ipynb

File Edit View Insert Runtime Tools Help

CODE TEXT CELL CELL COPY TO DRIVE RECONNECT EDIT

```
[ ]  import json
     from bs4 import BeautifulSoup
     import requests
     import pandas as pd
     import numpy as np
     import re
     pip install fake_useragent
     %matplotlib inline
     %pylab inline
```

```
Populating the interactive namespace from numpy and matplotlib
```

```
[ ]  !pip install fake_useragent
```

```
Collecting fake_useragent
  Downloading https://files.pythonhosted.org/packages/d1/79/af647635d6968e2deb57a208d309f6069d31cb138066d7e821e575112a80/fake-useragent-0.1.11.tar.gz
Building wheels for collected packages: fake-useragent
  Running setup.py bdist_wheel for fake-useragent ... done
  Stored in directory: /root/.cache/pip/wheels/5e/63/09/d1dc15179f175357d3f5c00cbffbac37f9e8690d80545143ff
Successfully built fake-useragent
Installing collected packages: fake-useragent
Successfully installed fake-useragent-0.1.11
```

```
from fake_useragent import UserAgent
UserAgent().chrome
```

```
Error occurred during loading data. Trying to use cache server https://fake-useragent.herokuapp.com/browsers/0.1.11
```

The issue

4 million companies

Website is blocking very quickly requesting

- Timer of random waiting from 1 to 20 seconds → 40 requests before being blocked
- How to unblock ?

Going around limits

- Change user agent
  - Fake_user agent

```
[ ]  from fake_useragent import UserAgent
     ua = UserAgent()
     user_agent_list  = ua.data_browsers['chrome']
     random.choice(user_agent_list)
```

```
'Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.2 Safari/537.36'
```

```
[ ]  user_agent_list
```

```
['Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2228.0 Safari/537.36',
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2227.1 Safari/537.36',
 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2227.0 Safari/537.36',
 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2227.0 Safari/537.36',
 'Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2226.0 Safari/537.36',
 'Mozilla/5.0 (Windows NT 6.4; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2225.0 Safari/537.36',
 'Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2225.0 Safari/537.36',
 'Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2224.3 Safari/537.36',
 'Mozilla/5.0 (Windows NT 10.0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.93 Safari/537.36',
 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/37.0.2062.124 Safari/537.36',
 'Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/37.0.2049.0 Safari/537.36',
```

```
def parse_companies_with_link(idx, startup_link, is_random=False):
    if is_random:
        sleep_time = random.randrange(1, 20)
        time.sleep(sleep_time)

    search_response = requests.get(startup_link, headers={'User-Agent': random.choice(user_agent_list)})
    soap = BeautifulSoup(search_response.content, 'html')
```

Going around limits

- Speed