

Python

Лекция 4: Парсинг данных. Регрессия

Отметься на портале!



План занятия

- Введение в веб
- Парсинг
- Парсинг данных
- Флask
- Введение в линейную регрессию
- Нелинейная регрессия

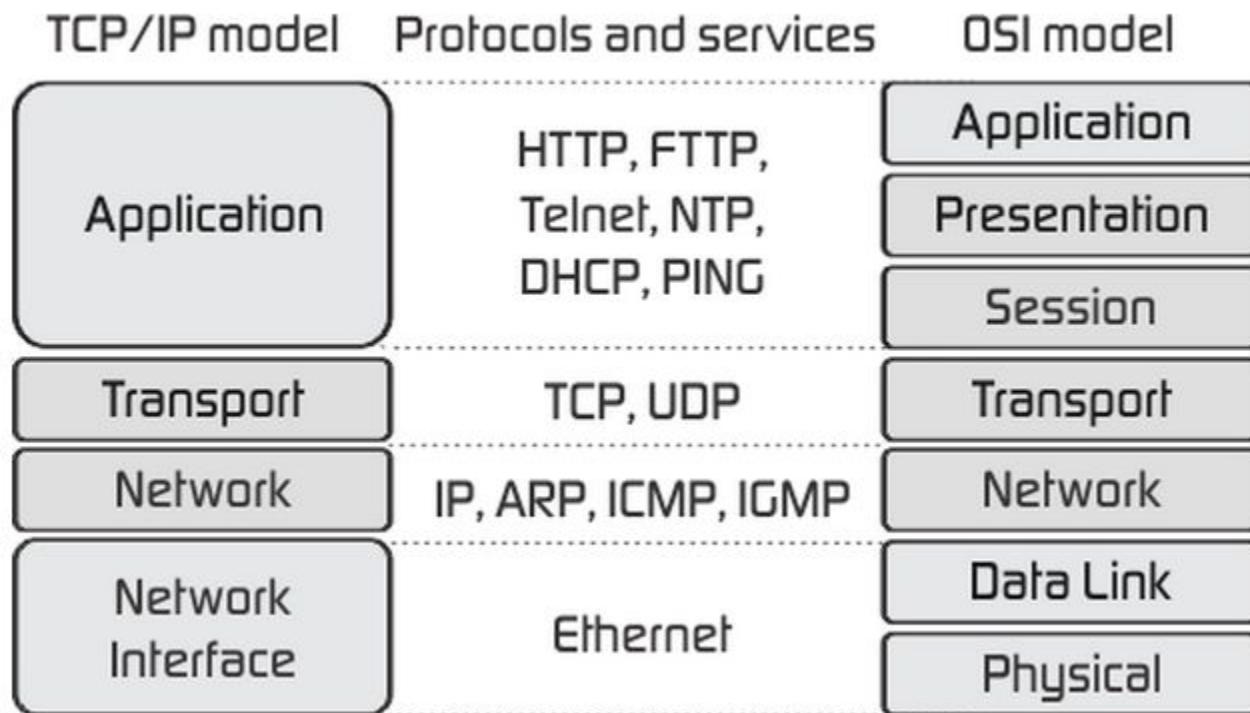
Подготовимся к лекции

1. Переходим в директорию вашего форка
2. Открываем в ней терминал
3. `git checkout master`
4. `git pull upstream master`
5. `git push -u origin master`

Сетевые протоколы

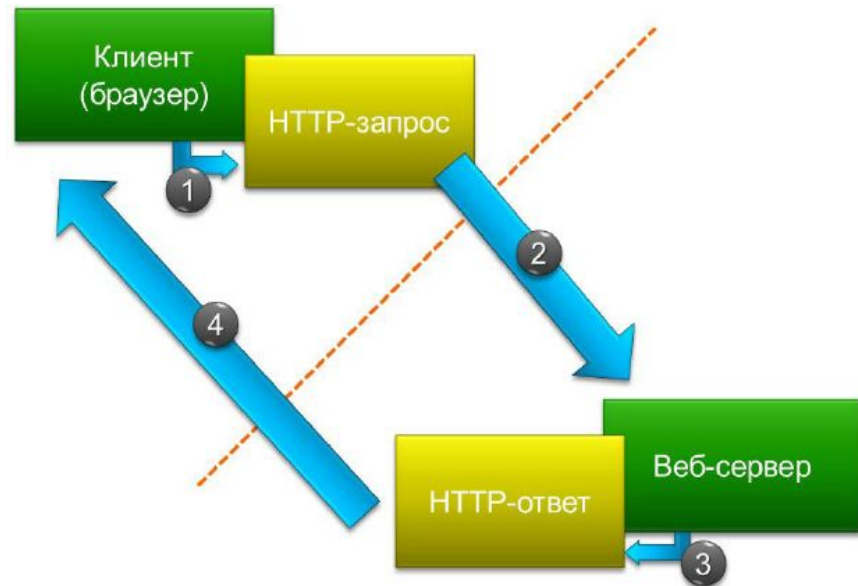
- Сеть -- совокупность устройств, подключенных друг к другу (физически или логически) и общающихся между собой
- Сетевой протокол -- набор правил и действий для “общения” >2 устройств, подключенных к сети
- Модель OSI
- Модель TCP/IP

Модель OSI & TCP/IP



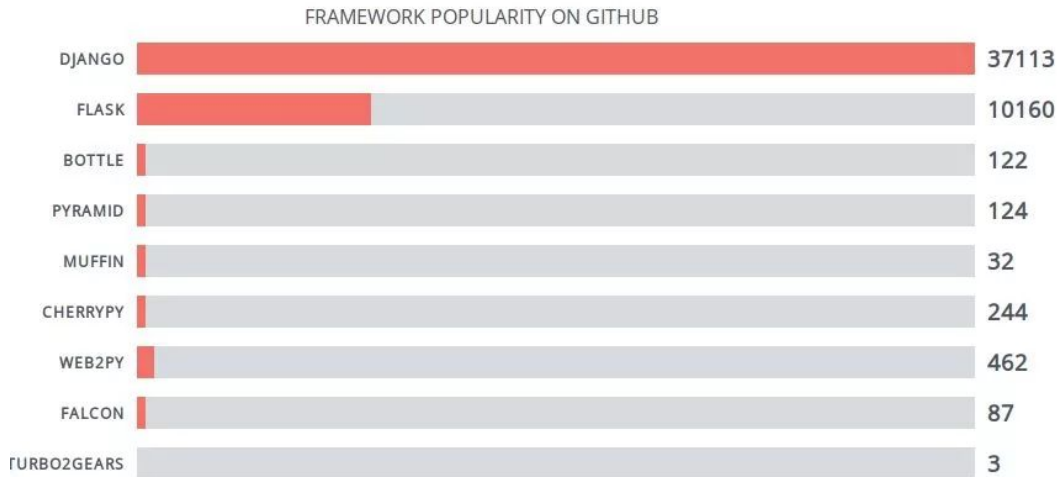
Протокол HTTP(S)

- HTTP (HyperText Transfer Protocol) -- протокол передачи гипертекста
- Клиент-серверное общение:
 - Клиент формирует (http) запрос
 - Сервер обрабатывает запрос и отвечает



Flask

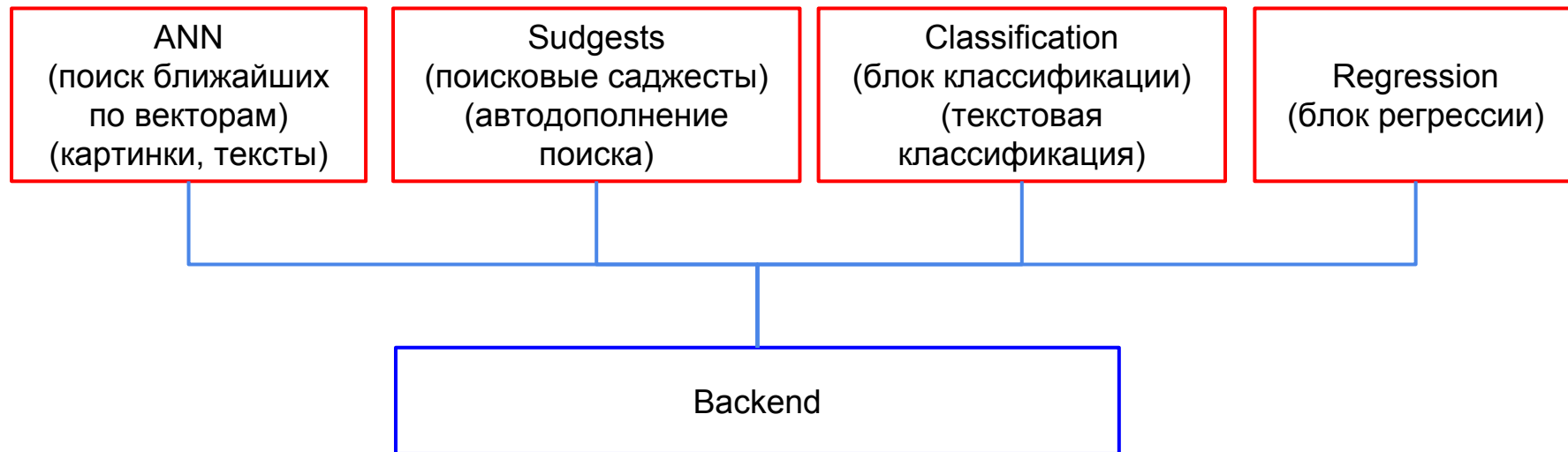
Микро фреймворк для
создания веб приложений



<http://flask.pocoo.org/docs/1.0/quickstart/#quickstart>

<http://flask.pocoo.org/docs/1.0/tutorial/>

Flask



Flask

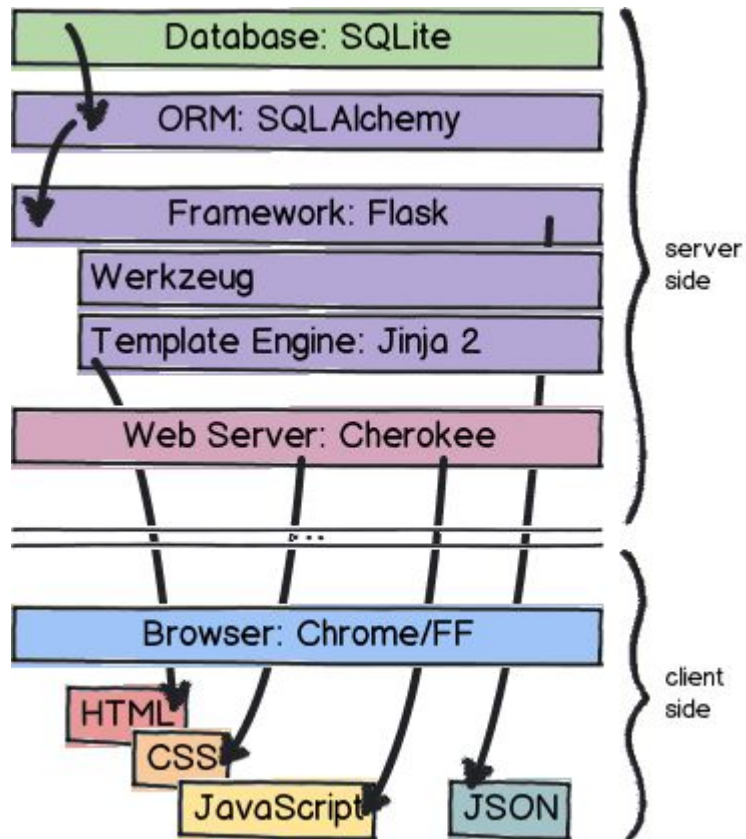
```
from flask import Flask
```

```
app = Flask(__name__) #app is the name of the object here
```

```
|  
from flask import Flask  
app = Flask(__name__)
```

```
@app.route("/")  
def hello():  
    return "Hello World!"
```

```
if __name__ == "__main__":  
    app.run()
```



Взгляд с другой стороны. Парсинг

Зачем вообще нам что-то парсить?



Алгоритм парсинга

- 1) Можно не парсить - не парсить
- 2) Есть возможность получить api-ключ - проще получить
- 3) Контент динамический?
 - а) да: Понять какой запрос идет на сервер
 - б) нет: Запрашивать страницу, понять структуру страницы
- 4) Есть ли защита от парсинга?
 - а) нет: Просто грузим страницы
 - б) да: подмена headers и user-agent, использование сетей прокси, selenium, эмуляция поведения человека

Можно не парсить - не парсить

1) Погуглить известные хранилища открытых данных

- a) <https://data.mos.ru/>
- b) <https://www.kaggle.com/>
- c) <https://docs.google.com/spreadsheets/d/1ZSLP1McnXv0FtOd9t7dMp3AfaiusvaGwWV0F9g2pbho/edit#gid=0>



Статический контент, защиты от парсинга нет.

[illegible]

Спарсили, и что делать?

Простейший случай. BeautifulSoup

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')
```

```
soup = BeautifulSoup('<b class="boldest">Extremely bold</b>', 'html.parser')
tag = soup.b
type(tag)
```

bs4.element.Tag

Тэг, у тега есть атрибуты

```
tag.attrs
```

```
{'class': ['boldest']}
```



Динамический контент

Динамический контент, защиты от парсинга нет.

[Пример сайта с динамическим контентом](#)

```
</div>
<div class="item-view-similar">
<div class="similar js-similar similar-column-4"
data-show-more-btn="1"> <div class="similar-inner similar-inner-hidden js-similar-inner"> <div class="similar-title">Похожие объявления</div>
```

Что делать?

1. Консоль разработчика
2. Найти запрос, по которому возвращается контент
3. Эмулировать запросы от клиента

vsLLM3bAOB_T3lhi7G4PkQupi1peCuDu5Rdf4psTYSITUZL-C...MLulhGoQAspkQZylzruJX...	200	document	pFr8KfFS3QOkLJ2lgk_CEHkRn0mwFolM...	20.0 KB	769 ms	
UFivLLM3bAOB_T3lhi7G4PkQupi1peCuDu5Rdf4psTYSITUZ...LJeptCJUyooMHfS8hmaG...	200	document	pFr8KfFS3QOkLJ2lgk_CEHkRn0mwFolM...	19.4 KB	2.44 s	
uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5H90...bJfRrAY4XslwTeyASlbOT...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.88 s	
Fr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5...ZJfBqD5HqouUFasQtmC...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5H90...JPvE5CpkMulsBay8qkaZ...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.88 s	
Fr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5...TOOdIBpMzJgWYCW9gr...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.88 s	
uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5H90...YOoD6S5ABtoEaaTE4La...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
pFr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb...VKfhrBoUar44IayQ7kaOaR...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5H9...biOT4Ap8JrI4YZjkgaedX...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
hpFr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVg...fL-t4G5KkQlQZzwtg7ydvN...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.86 s	
hpFr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVg...aNP58ColAwg8BYCQhgaCZ...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
Fr4uQfVWHbikHP1orJ5m4HkAijxgSpIwAUlJ1TdZYSEC_1BVgb5...aNP58ColAwg8BYCQhgaCZ...	200	fetch	r8KfFS3eLEaT2lok62BHxnm0m09vNma...	172 B	1.87 s	
data?_rmd=1540242082537&referrer=https%3A%2F%2Fwww...u%2Fmoskva%2Ftelefony...	200	fetch	VM1229:r8KfFS3QOkLJ2lgk_CEHkRn0...	1.1 KB	579 ms	
1?page-ref=https%3A%2F%2Fwww.avito.ru%2Fmoskva%2F...0%89%D1%82%D0%B5%2...	200	gif	tao.js:283	535 B	403 ms	
fe?data={%22hostname%22:%22www.avito.ru%22,%22brow...%22:2312,%22onLoad%22:...	200	gif	fe_metric.min.js:1	284 B	1.79 s	
?random=1208658271&cv=9&fst=*&num=1&value=0&label=...mpgE&random=873462467...	200	gif	www.google.com/	87 B	130 ms	
?random=1208658271&cv=9&fst=*&num=1&value=0&label=...&ocp_id=ojrOW-njB5KeYqq...	302	gif	?random=1540242082100&cv=9&fst=1...	685 B	707 ms	
?random=1208658271&cv=9&fst=*&num=1&value=0&label=...YeKqY1mpgE&random=873...	302	gif	googleads.g.doubleclick.net/	808 B	230 ms	
setuid?entity=430&code=E56DF7AC7BE88511	200	gif	an.yandex.ru/mapid/appnexus/	2.2 KB	382 ms	
YAgA?time=1540242082.763	200	gif	stats.mos.ru/gc/vnd/	328 B	745 ms	

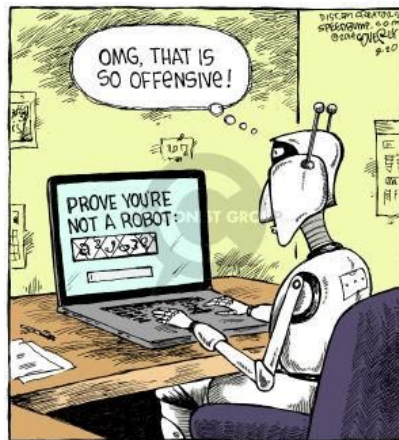
Защита от парсинга

Способы:

- капча
- бан
- левые статические страницы

Что делать?

1. Подстановка user-agent, headers
2. Таймауты
3. Прокси, тор
4. Парсинг из точек с большим лимитом (общага =))



Accept-Language
User-Agent
Accept-Encoding
Host
DNT
Connection
Cache-Control
Cookie

Debugger	Network	UI Responsiveness	Profiler	Memory
<input type="text"/>				
http://www.bing.com/				
Response headers				
Response body				
Cookies				
Initiator				
Timings				
Value				
GET / HTTP/1.1				
text/html, application/xhtml+xml, */*				
en-US				
Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko				
gzip, deflate				
www.bing.com				
1				
Keep-Alive				
no-cache				
SRCHUID=V=2&GUID=0CD0863343104DC781CA455E13B79674; MJUIDB=1D...				

Тяжелая артиллерия. Selenium

Более сложный случай - сильная защита, контент типа саджестов, который сложно эмулировать запросами итд итп

Что делать?

Заставим компьютер быть похожим на человека

XPath:

```
/wikimedia/projects/project/editions/*[2]
```

XML document:

```
<?xml version="1.0" encoding="utf-8"?>
<wikimedia>
  <projects>
    <project name="Wikipedia" launch="2001-01-05">
      <editions>
        <edition language="English">en.wikipedia.org</edition>
        <edition language="German">de.wikipedia.org</edition>
        <edition language="French">fr.wikipedia.org</edition>
        <edition language="Polish">pl.wikipedia.org</edition>
      </editions>
    </project>
    <project name="Wiktionary" launch="2002-12-12">
      <editions>
        <edition language="English">en.wiktionary.org</edition>
        <edition language="French">fr.wiktionary.org</edition>
        <edition language="Vietnamese">vi.wiktionary.org</edition>
        <edition language="Trukish">tr.wiktionary.org</edition>
      </editions>
    </project>
  </projects>
</wikimedia>
```



Давайте поиграем

Хочется с <https://fred.stlouisfed.org/> спарсить все временные ряды, относящиеся к США.

- 1) Какой тип контента?
- 2) Как будем действовать?

Вернемся к yahoo.finance

- 1) Как можно по-другому спарсить данные?
- 2) Давайте попробуем сделать это



Домашнее задание

Подготовка данных для проекта. С каждого PR с парсером (по парсеру на человека). Ветка homework_04

Источники

[Документа супа](#)

[Flask мануал](#)

[Объемный обзор фласка](#)

[Хаброучебник фласка](#)

[Краткий и емкий мануал по requests](#)