

Парсер интернет- магазинов

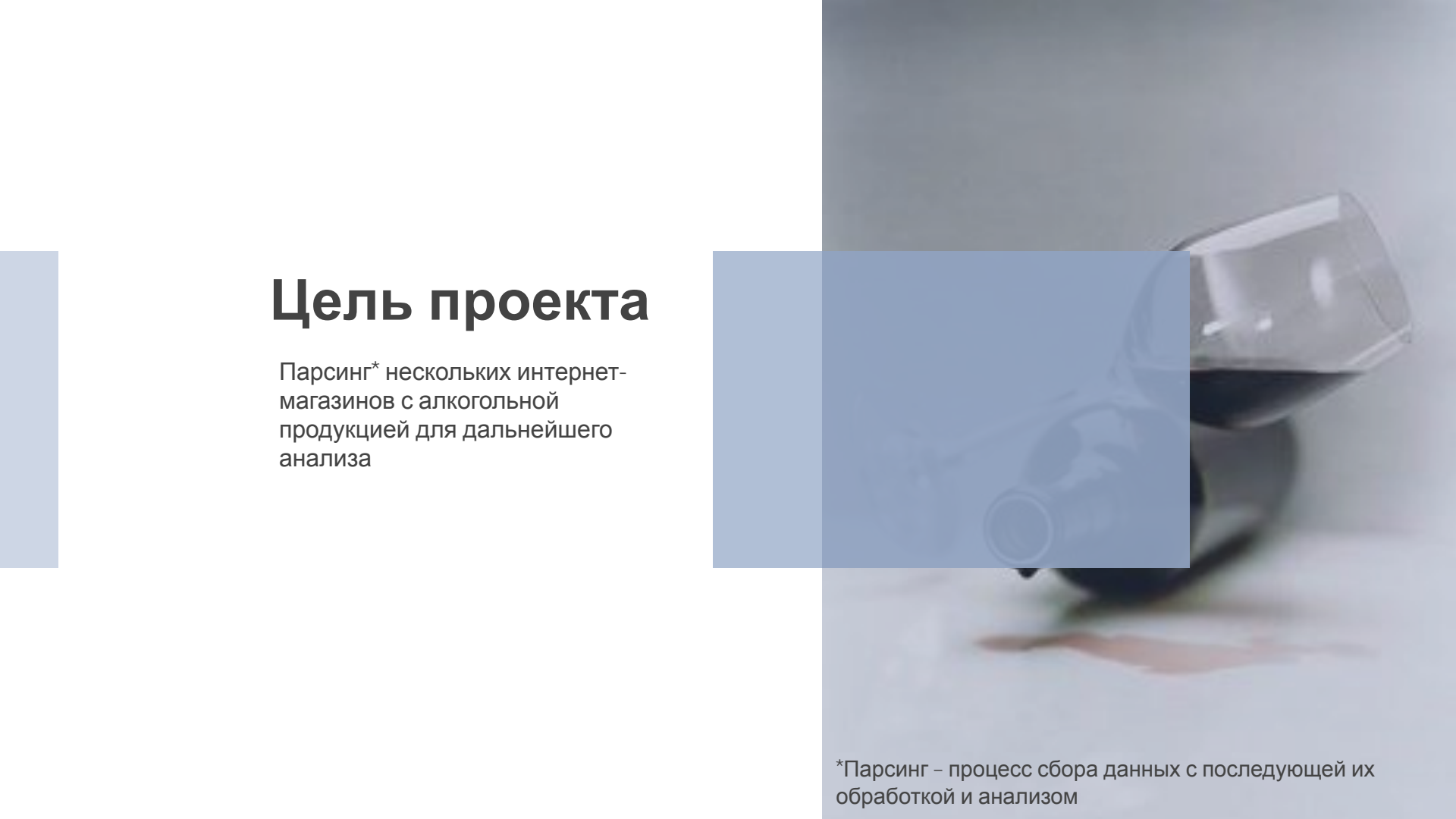
Ревенко Дарья, Заболотская
Алина

<https://github.com/arzabolotskaya/LCOPARSE>



Цель проекта

Парсинг* нескольких интернет-магазинов с алкогольной продукцией для дальнейшего анализа



*Парсинг – процесс сбора данных с последующей их обработкой и анализом

Наши задачи



Узнать, что такое парсинг

Изучить понятие парсинга, освоить BeautifulSoup и Selenium

Спарсить 9 сайтов

Создать шаблоны для парсинга сайтов и парсить сайты раз в несколько недель

Оценить динамику цен

Провести аналитику и сравнить цены на алкоголь в универсальных и специальных магазинах, в одном магазине на протяжении двух месяцев

Как мы это делаем

Этап 1

```
def get_html(url, params = None):
```

```
    r = requests.get(url, headers = HEADERS, params = params)
```

```
    return r
```

Достаем html-код со
страницы

Этап 2

Достаем нужный
нам контент

```
    alkogol.append({  
        'category': 'Крепкие напитки',  
        'title': item.find('div', class_ = "b-product__title-block").get_text(strip = True),  
        'link': HOST + item.find('a', class_ = "b-product__title js-list-prod-open").get('href'),  
        'price': item.find('div', class_ = "b-product__price").get('content').replace('\xa0', ''),  
        'special_offer': special_offer  
    })
```

Этап 3

```
def save_file(items, path):
```

```
    with open (path, 'w', newline = '') as f:
```

```
        writer = csv.writer(f, delimiter = ';')
```

```
        writer.writerow(['Категория', 'Название', 'Ссылка', 'Цена', 'Акции'])
```

```
        for item in items:
```

```
            writer.writerow([item['category'], item['title'], item['link'], item['price']])
```

Сохраняем в файл csv-
формата для дальнейшей
работы

Как это реализовано Селениумом

Этап 1

Для парсинга
используем драйвер

```
def main():  
    driver = webdriver.Chrome()  
    parser = ProgHubParser(driver)  
    all_alkogol = parser.parse()  
  
    save_file(all_alkogol, PATH)
```

Этап 2

Достаем нужный
нам контент

```
try:  
    content_name_elm = card.find_element_by_class_name("product_item_name").find_element_by_tag_name("a")  
    content.name = content_name_elm.text  
except NoSuchElementException:  
    print("product name missing")
```

Этап 3

Сохраняем в файл csv-
формата для дальнейшей
работы

```
def save_file(items, path):  
    with open(path, 'w', newline = '') as f:  
        writer = csv.writer(f, delimiter = ';')  
        writer.writerow(['Категория', 'Название', 'Ссылка', 'Цена', 'Акции'])  
        for item in items:  
            writer.writerow([item['category'], item['title'], item['link'], item['price']])
```

Главная вещь в программе

```
def get_content(html):
    soup = BeautifulSoup(html, 'html.parser')
    items = soup.find_all("div", class_ = "b-grid_item")
    alkogol = []
    for item in items:
        offer = item.find('span', class_="b-product-tag")
        if offer:
            special_offer = offer.get_text(strip = True).replace('\n', '')
        else:
            special_offer = "На данный товар нет акций"
        alkogol.append({
            'category': 'Крепкие напитки',
            'title': item.find('div', class_ = "b-product_title-block").get_text(strip = True),
            'link': HOST + item.find('a', class_="b-product_title js-list-prod-open").get('href'),
            'price': item.find('div', class_ = "b-product_price").get('content').replace('\xa0', ''),
            'special_offer': special_offer
        })
    return alkogol
```

Данные в цифрах

9

Количество магазинов

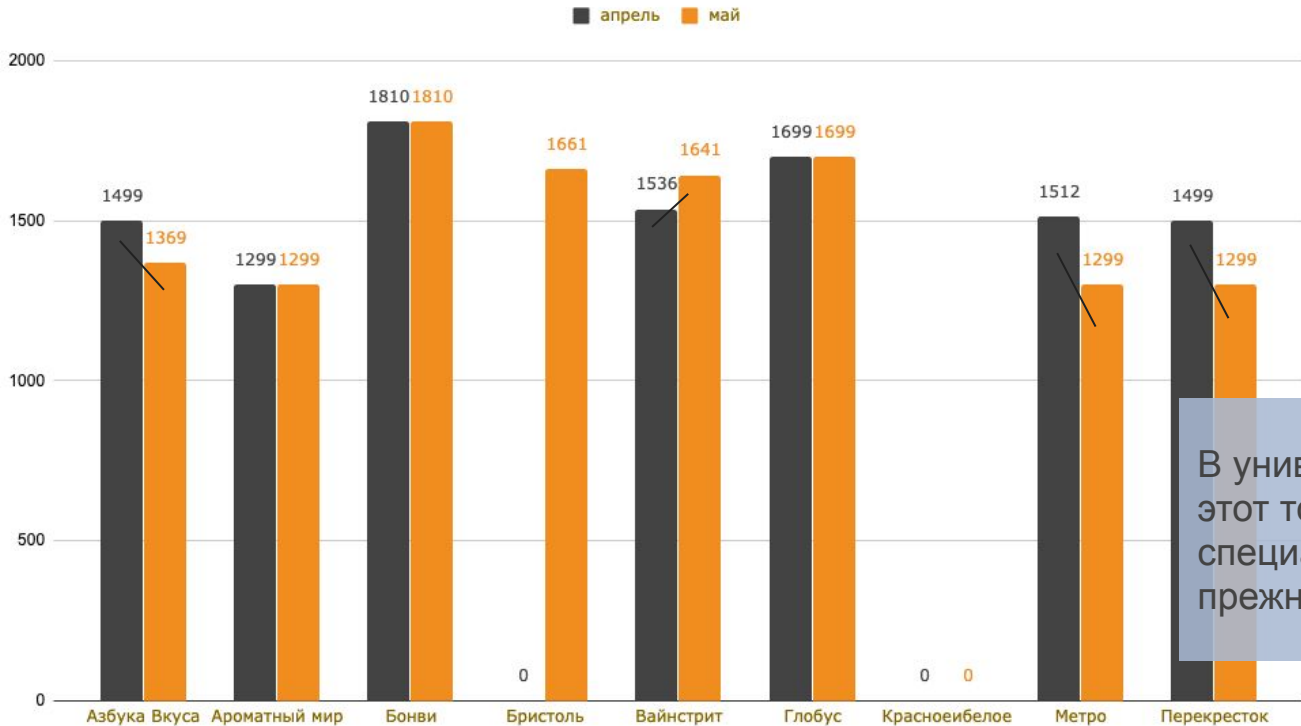
>60

Дней работы над
проектом

>40000

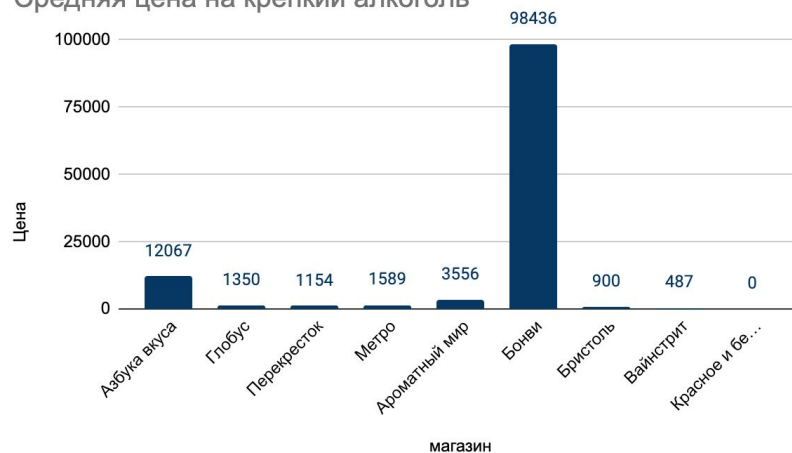
Количество товаров во
всех магазинах

Цена на текилу "Olmeca" Blanco, 0.7 л



В универсальных магазинах цена на этот товар почти везде снизилась, в специальных магазинах осталась прежней или повысилась

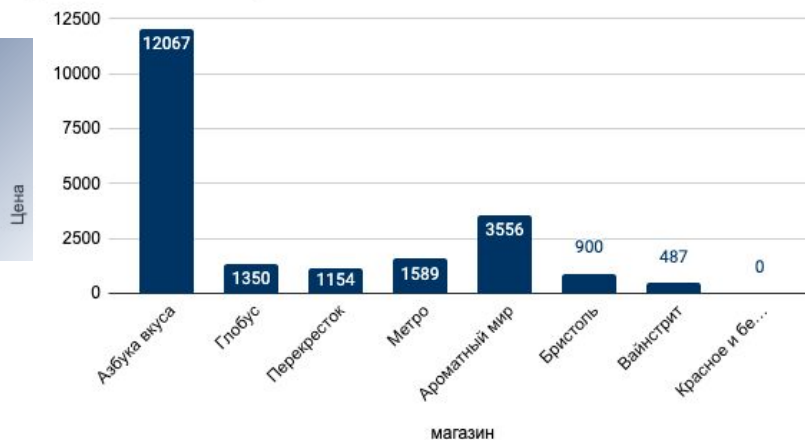
Средняя цена на крепкий алкоголь



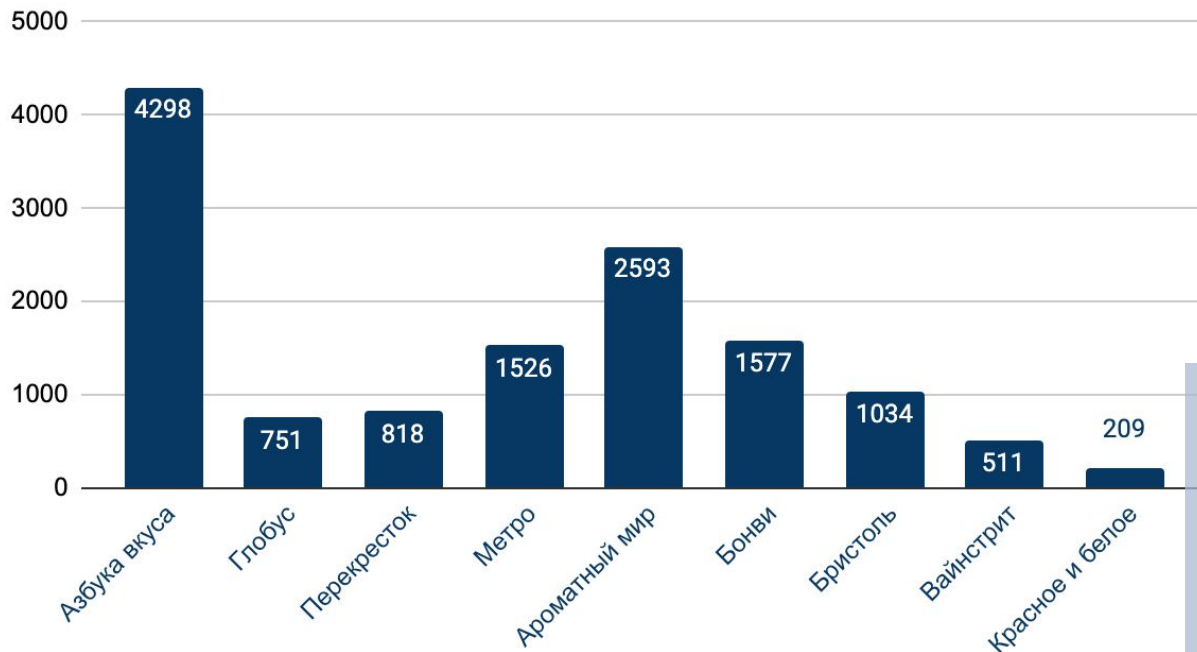
Аналитика

Если убрать из выборки Бонви, где средняя цена на крепкий алкоголь сильно выше, чем в остальных магазинах, можно убедиться, что самое большое среднее значение у азбуки вкуса.

Средняя цена на крепкий алкоголь

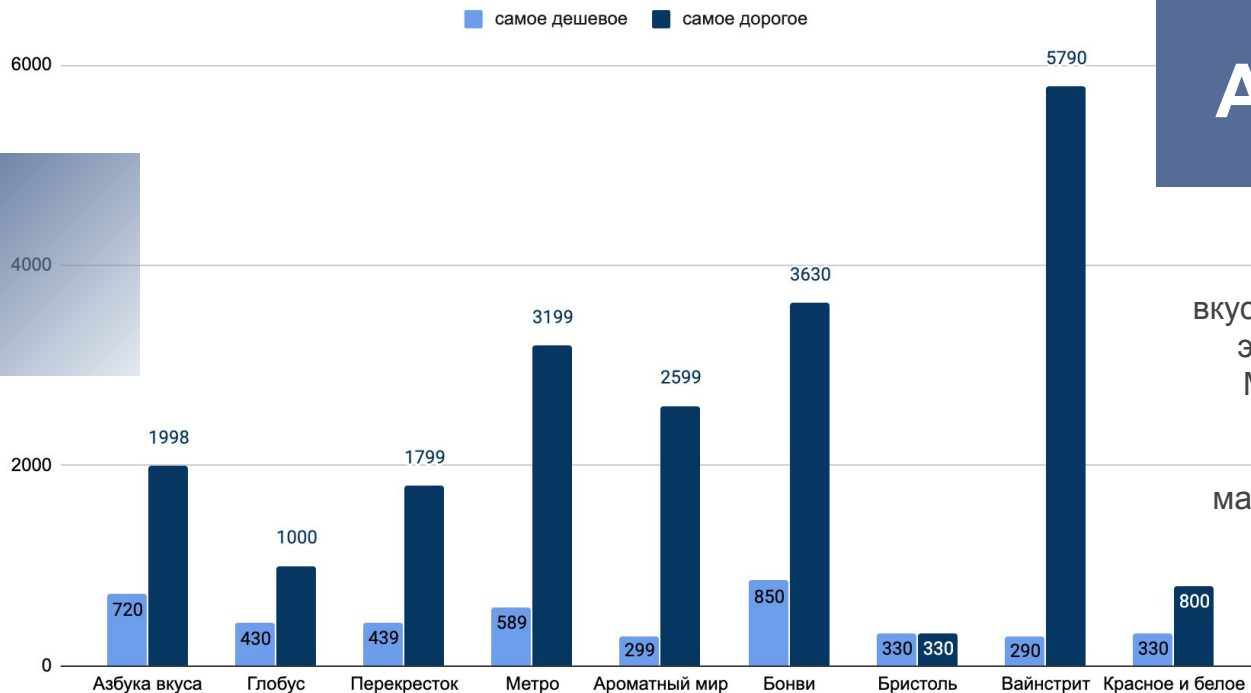


Средняя цена вина во всех магазинах



Средняя цена на вино в азбуке вкуса как минимум в 1,5 раза выше, чем в других магазинах.

самое дешевое и самое дорогое из pinot grigio



Аналитика

Во всех магазинах, кроме “азбуки вкуса” и “бонви” минимальная цена на это вино примерно 300-500 рублей. Максимальная цена колеблется от 1000 до почти 6000 рублей.

Самое дорогое вино этого вида в магазине “красное и белое” дешевле самого дешевого вина из “бонви”.

Самый большой “разброс” цен - Вайнстрит

1550

1500

1450

1400

1350

1300

1250

1200

1150

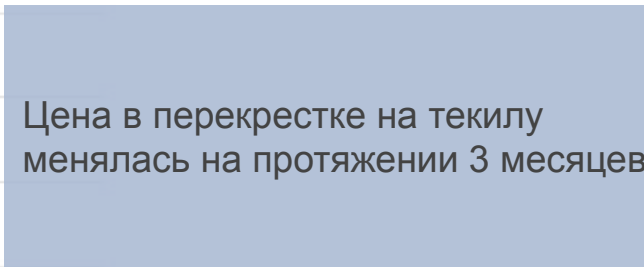
март

апрель

май

— Цена на текилу Olmeca Blanco в Перекрестке

Цена в перекрестке на текилу
менялась на протяжении 3 месяцев





ПАГИНАЦИЯ

Отсутствие классической пагинации на сайтах(1,2;3 и т д) затрудняло процесс парсинга



БАНЫ

Периодически было сложно вытащить нужный контент из-за блокировок

Трудности



АНАЛИЗ

На разных сайтах разные названия одного и того же товара → невозможность автоматического анализа



SELENIUM

Для парсинга некоторых сайтов потребовалось изучить selenium

Выводы

- Азбука вкуса и Бонви оказались магазинами с самыми большими средними ценами на алкоголь
- Не во всех интернет магазинах можно наблюдать динамику цен.
- Парсер является эффективным инструментом автоматизации сбора большого количества данных с веб-страниц, хотя имеет ряд проблем
- Можно наблюдать значительные различия в ассортименте разных интернет магазинов: как в разнообразии представленной продукции, так и в ее ценовом диапазоне.



Автоматизация

Придумать способ
унификации названий

Изображения

Освоить парсинг
изображений –
искать сходство по
картинке



Для ускорения
обработки данных

Многопоточность

Развитие проекта