

Твердая цифра

Новый источник данных по потребительским ценам

Александр Исаков Родион Латыпов Андрей Репин
Егор Постолит Алексей Евсеев Елена Синельникова-Мурылева

22 декабря 2020 г.

Содержание

1. природа данные и цель

Содержание

1. природа данные и цель
2. международный опыт статистических институтов

Содержание

1. природа данные и цель
2. международный опыт статистических институтов
3. набор технологий для мониторинга цен

Содержание

1. природа данные и цель
2. международный опыт статистических институтов
3. набор технологий для мониторинга цен
4. масштаб собираемых данных

Содержание

1. природа данные и цель
2. международный опыт статистических институтов
3. набор технологий для мониторинга цен
4. масштаб собираемых данных
5. текущая повестка нашей работы

Природа данных

- ▶ подход: вэб-скрэйпинг для мониторинга потребительских цен

Природа данных

- ▶ **подход:** вэб-скрэйпинг для мониторинга потребительских цен
- ▶ **периметр:** потребительские цены продовольственных, непродовольственных товаров и услуг

Природа данных

- ▶ **подход:** вэб-скрэйпинг для мониторинга потребительских цен
- ▶ **периметр:** потребительские цены продовольственных, непродовольственных товаров и услуг
- ▶ **цель:** создать открытый источник гранулярных исторических данных по ценам в квази-реальном времени для исследователей

Обзор литературы

вэб-скрэйпинг - не относится к "новым" или
"экспериментальным" подходам к мониторингу цен

Мы можем опереться на широкий международный опыт:

1. национальных статистических институтов: ([Eurostat, 2014](#)), ([Istat, 2015](#)), ([ONS UK, 2017](#)), ([BLS, 2013](#)), ([Statistics New Zealand, 2015](#)) и др.

Обзор литературы

вэб-скрэйпинг - не относится к "новым" или
"экспериментальным" подходам к мониторингу цен

Мы можем опереться на широкий международный опыт:

1. национальных статистических институтов: ([Eurostat, 2014](#)), ([Istat, 2015](#)), ([ONS UK, 2017](#)), ([BLS, 2013](#)), ([Statistics New Zealand, 2015](#)) и др.
2. центральных банков: ([Lünnemann & Wintr, 2006](#)) в ЕЦБ, ([Lazyan et. al, 2017](#)) в ЦБА, ([Hull et al., 2017](#)) в Банке Швеции, ([Macias & Stelmasiak, 2019](#)) в Банке Польши, ([Ellul, 2019](#)) в Банке Мальты и д.р.

Обзор литературы

вэб-скрэйпинг - не относится к "новым" или
"экспериментальным" подходам к мониторингу цен

Мы можем опереться на широкий международный опыт:

1. национальных статистических институтов: ([Eurostat, 2014](#)), ([Istat, 2015](#)), ([ONS UK, 2017](#)), ([BLS, 2013](#)), ([Statistics New Zealand, 2015](#)) и др.
2. центральных банков: ([Lünnemann & Wintr, 2006](#)) в ЕЦБ, ([Lazyan et. al, 2017](#)) в ЦБА, ([Hull et al., 2017](#)) в Банке Швеции, ([Macias & Stelmasiak, 2019](#)) в Банке Польши, ([Ellul, 2019](#)) в Банке Мальты и д.р.
3. академических исследователей: наиболее известен The Billion Prices Project Alberto Cavallo из MIT и NBER, см ([Cavallo & Rigobon, 2012](#))

Обзор литературы

вэб-скрэйпинг - не относится к "новым" или
"экспериментальным" подходам к мониторингу цен

Мы можем опереться на широкий международный опыт:

1. национальных статистических институтов: ([Eurostat, 2014](#)), ([Istat, 2015](#)), ([ONS UK, 2017](#)), ([BLS, 2013](#)), ([Statistics New Zealand, 2015](#)) и др.
2. центральных банков: ([Lünnemann & Wintr, 2006](#)) в ЕЦБ, ([Lazyan et. al, 2017](#)) в ЦБА, ([Hull et al., 2017](#)) в Банке Швеции, ([Macias & Stelmasiak, 2019](#)) в Банке Польши, ([Ellul, 2019](#)) в Банке Мальты и д.р.
3. академических исследователей: наиболее известен The Billion Prices Project Alberto Cavallo из MIT и NBER, см ([Cavallo & Rigobon, 2012](#))
4. частного сектора: любая розничная сеть, агрегатор цен (например, [Yandex.Market](#) в России, [Trundler](#) в ЮАР) и др.

Набор технологий

Три базовых процесса: 1) сбор данных, и 2) хранение и доступ к данным, 3) совместная разработка.

Сбор данных:

1. Python + Selenium + requests

Набор технологий

Три базовых процесса: 1) сбор данных, и 2) хранение и доступ к данным, 3) совместная разработка.

Сбор данных:

1. Python + Selenium + requests
2. ad-hoc решения: решение CAPTCHA, подтверждение по SMS

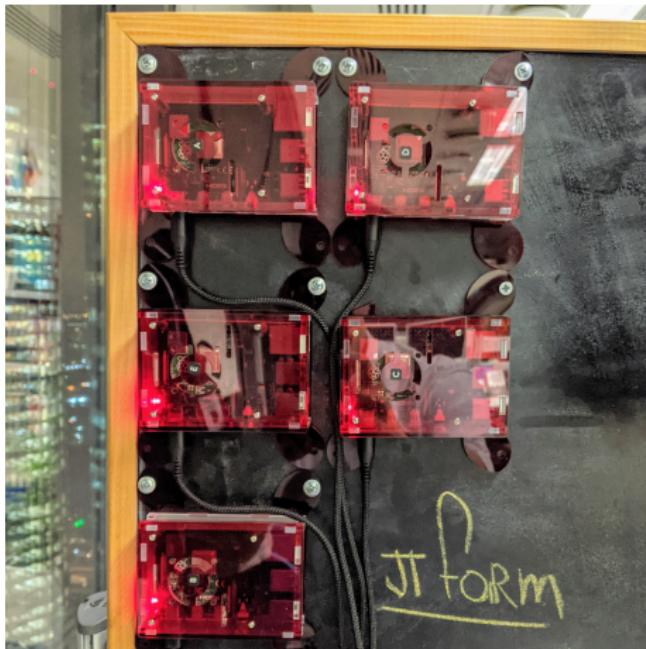
Набор технологий

Три базовых процесса: 1) сбор данных, и 2) хранение и доступ к данным, 3) совместная разработка.

Сбор данных:

1. Python + Selenium + requests
2. ad-hoc решения: решение CAPTCHA, подтверждение по SMS
3. инфраструктура: отдельный ПК → ферма Raspberry Pi → Docker

Сбор данных - первый опыт



Набор технологий

Хранение и доступ к данным:

1. Хранение данные: MS SQL Server внутри организации

Набор технологий

Хранение и доступ к данным:

1. Хранение данные: MS SQL Server внутри организации
2. Распространение данных: облачный PostgreSQL + WEB API

Набор технологий

Хранение и доступ к данным:

1. Хранение данные: MS SQL Server внутри организации
2. Распространение данных: облачный PostgreSQL + WEB API
3. Важно: WEB API позволяет и получать данные, и совместно накапливать

Хранение и доступ к данным

Интерактивное руководство по работе API на Google Colab

Линолеум (RosstatID = 7411). Как выгрузить таблицу соответствия между товарной категорией и RosstatID, см. внизу

Достанем всю информацию из таблицы rtpi_price_page со всеми линолеумами. Как ограничивать запрос по датам, см. наверху

```
rosstat_id = 7411

request_url = f"http://164.90.194.12:8080/rtpi_price_page?select=*&rosstat_id={rosstat_id}"

response = requests.get(request_url, headers = request_headers)
response.json()

[{'contributor_id': 1,
 'date_add': '2020-06-29T14:38:15.81',
 'date_last_crawl': '2020-11-11T16:14:39.267',
 'date_last_in_stock': '2020-11-11T16:14:39.267',
 'price_name': 'https://leroymerlin.ru/product/linoleum-dub-sevilskiy-31-klass-3-m-18618811/',
 'price_url': 'https://leroymerlin.ru/product/linoleum-dub-sevilskiy-31-klass-3-m-18618811/'},
 {'contributor_id': 1,
 'date_add': '2020-06-29T14:38:15.81',
 'date_last_crawl': '2020-11-11T16:14:57.093',
 'date_last_in_stock': '2020-11-11T16:14:57.093',
 'price_name': 'https://leroymerlin.ru/product/linoleum-tempo-dub-antik-33-klass-4-m-18618538/',
 'price_url': 'https://leroymerlin.ru/product/linoleum-tempo-dub-antik-33-klass-4-m-18618538/'},
 {'contributor_id': 1,
 'date_add': '2020-06-30T20:36:02.873',
 'date_last_crawl': '2020-10-20T15:47:23.99',
 'date_last_in_stock': '2020-10-20T15:47:23.99',
 'price_name': 'https://leroymerlin.ru/product/linoleum-dub-dymchatyy-22-klass-3-5-m-82361168/',
 'price_url': 'https://leroymerlin.ru/product/linoleum-dub-dymchatyy-22-klass-3-5-m-82361168/'},
```

Совместная разработка

- ▶ Наиболее "дорогая" часть: разработка скрэйперов

Совместная разработка

- ▶ Наиболее "дорогая" часть: разработка скрэйперов
- ▶ "Героические" индивидуальные проекты - не выживают или остаются карликовыми

Совместная разработка

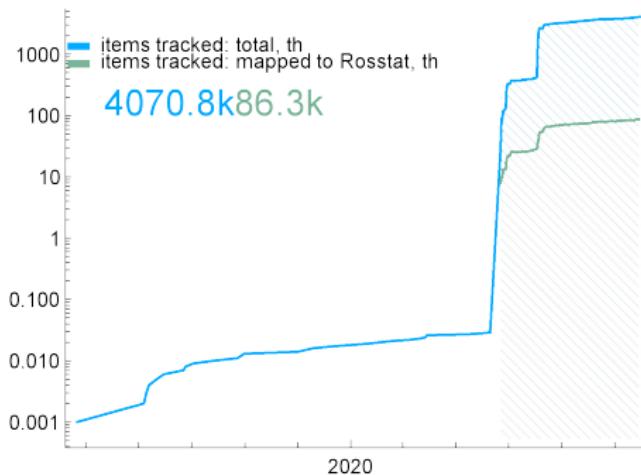
- ▶ Наиболее "дорогая" часть: разработка скрэйперов
- ▶ "Героические" индивидуальные проекты - не выживают или остаются карликовыми
- ▶ Кооперация: GitHub для разработки + WEB API для сбора

«Индивидуальный» подход



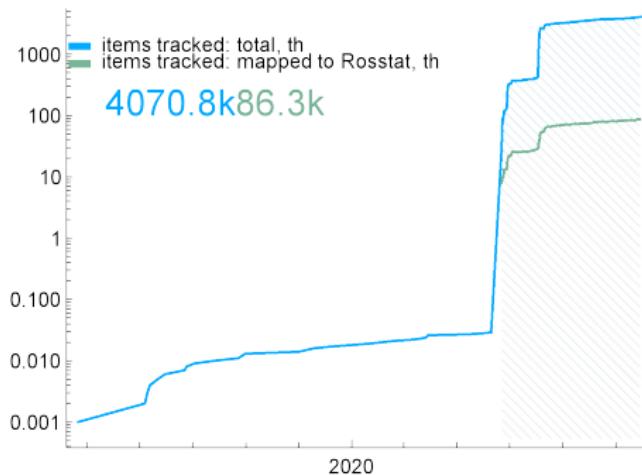
Масштаб

- ▶ 4.1M уникальных товаров и услуг



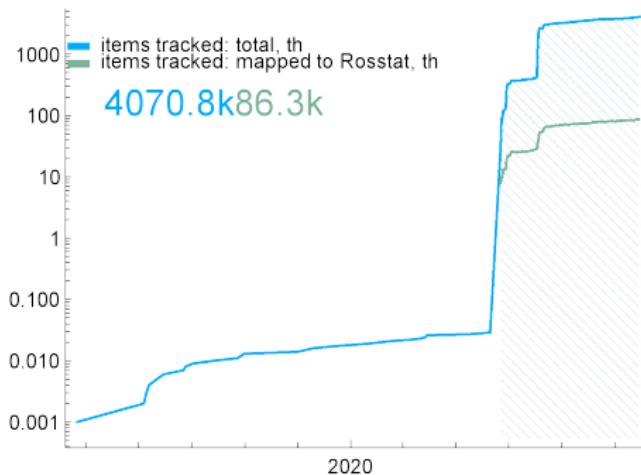
Масштаб

- ▶ 4.1М уникальных товаров и услуг
- ▶ 86 тыс. размечено в классификатор Росстата



Масштаб

- ▶ 4.1M уникальных товаров и услуг
- ▶ 86 тыс. размечено в классификатор Росстата
- ▶ частота наблюдения: от ежедневной до 3 часовой



Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.
- ▶ **StockStatus** - статус товара (в наличии/отсутствует)

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.
- ▶ **StockStatus** - статус товара (в наличии/отсутствует)
- ▶ **CurrentPrice** - возможная цена покупки/текущая цена

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.
- ▶ **StockStatus** - статус товара (в наличии/отсутствует)
- ▶ **CurrentPrice** - возможная цена покупки/текущая цена
- ▶ **CrossedPrice** - прежняя цена, обычно до скидки

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.
- ▶ **StockStatus** - статус товара (в наличии/отсутствует)
- ▶ **CurrentPrice** - возможная цена покупки/текущая цена
- ▶ **CrossedPrice** - прежняя цена, обычно до скидки
- ▶ **ProductName** - текстовое название товара

Состав отслеживаемых данных

Единицей наблюдения является отдельная цена в отдельном магазине: товар Т в магазине М.

- ▶ **PriceName** - уникальное название товара: обычно URL страницы товара в магазине, а когда цены данные "плоской страницей"/меню, то представляет собой сумму названия
- ▶ **PriceURL** - URL адрес цены/прайс-листа.
- ▶ **StockStatus** - статус товара (в наличии/отсутствует)
- ▶ **CurrentPrice** - возможная цена покупки/текущая цена
- ▶ **CrossedPrice** - прежняя цена, обычно до скидки
- ▶ **ProductName** - текстовое название товара
- ▶ Прочее: дата первого наблюдения, последнего наблюдения, последнего зарегистрированного наличия товара, идентификаторы Росстата, источника данных, поставщика наблюдения и т.п.

Альтернативные технологии сбора цен

- ▶ Иные технологии включают: данные ККТ, карточных расходов, стандартные методы наблюдения за ценами

Альтернативные технологии сбора цен

- ▶ Иные технологии включают: данные ККТ, карточных расходов, стандартные методы наблюдения за ценами
- ▶ Каждая из них имеет свои существенные сильные стороны и ограничения

Альтернативные технологии сбора цен

- ▶ Иные технологии включают: данные ККТ, карточных расходов, стандартные методы наблюдения за ценами
- ▶ Каждая из них имеет свои существенные сильные стороны и ограничения
- ▶ Вэб-скрэйпинг может быть сильной компонентой набора инструментов наблюдения за ценами

Особенности вэб-скрэйпинга

- ▶ **Низкая стоимость:** стоимость наблюдений ниже, чем у ККТ и стандартных методов из-за более простых технологий

Особенности вэб-скрэйпинга

- ▶ **Низкая стоимость:** стоимость наблюдений ниже, чем у ККТ и стандартных методов из-за более простых технологий
- ▶ **Высокая гранулярность:** позволяет более точно идентифицировать позицию, чем данные ККТ или карточных расходов, которые часто дают представление лишь об общей категории, МСС коде

Особенности вэб-скрэйпинга

- ▶ **Низкая стоимость:** стоимость наблюдений ниже, чем у ККТ и стандартных методов из-за более простых технологий
- ▶ **Высокая гранулярность:** позволяет более точно идентифицировать позицию, чем данные ККТ или карточных расходов, которые часто дают представление лишь об общей категории, МСС коде
- ▶ **Открытость:** выставленные цены являются публичной информацией, не требующей специальных технологий защиты персональных данных

Особенности вэб-скрэйпинга

- ▶ **Низкая стоимость:** стоимость наблюдений ниже, чем у ККТ и стандартных методов из-за более простых технологий
- ▶ **Высокая гранулярность:** позволяет более точно идентифицировать позицию, чем данные ККТ или карточных расходов, которые часто дают представление лишь об общей категории, МСС коде
- ▶ **Открытость:** выставленные цены являются публичной информацией, не требующей специальных технологий защиты персональных данных
- ▶ **Периметр покрытия:** данные ККТ и карточные данные дают информацию по фактическим сделкам, тогда как вэб-скрэйпинг дает все ценники - что ближе к традиционным методам

Особенности вэб-скрэйпинга

Хочу обратить внимание:

- ▶ **мы не говорим об индексах** - наша работа в создании источника сырых данных

Особенности вэб-скрэйпинга

Хочу обратить внимание:

- ▶ **мы не говорим об индексах** - наша работа в создании источника сырых данных
- ▶ квалифицированный специалист может использовать свое суждение для построения индексов на их основе

Особенности вэб-скрэйпинга

Хочу обратить внимание:

- ▶ **мы не говорим об индексах** - наша работа в создании источника сырых данных
- ▶ квалифицированный специалист может использовать свое суждение для построения индексов на их основе
- ▶ **мы не говорим о федеральной репрезентативности** - сегодня мы стремимся к полному покрытию тестового региона

Заключение - открытые задачи

- ▶ мы ищем коллег - которые видят потребность в доступной детальной информации по потребительским ценам;

Заключение – открытые задачи

- ▶ **мы ищем коллег** – которые видят потребность в доступной детальной информации по потребительским ценам;
- ▶ **мы ищем аналитиков** – которые понимают, что наши данные открывают возможности для получения нового знания о ценах;

Заключение - открытые задачи

- ▶ **мы ищем коллег** - которые видят потребность в доступной детальной информации по потребительским ценам;
- ▶ **мы ищем аналитиков** - которые понимают, что наши данные открывают возможности для получения нового знания о ценах;
- ▶ **мы ищем конструктивную критику** - любые замечания и идеи по нашему подходу.

Связь

- ▶ ai@vtbcapital.com

Связь

- ▶ ai@vtbcapital.com
- ▶ <https://t.me/xiskv>