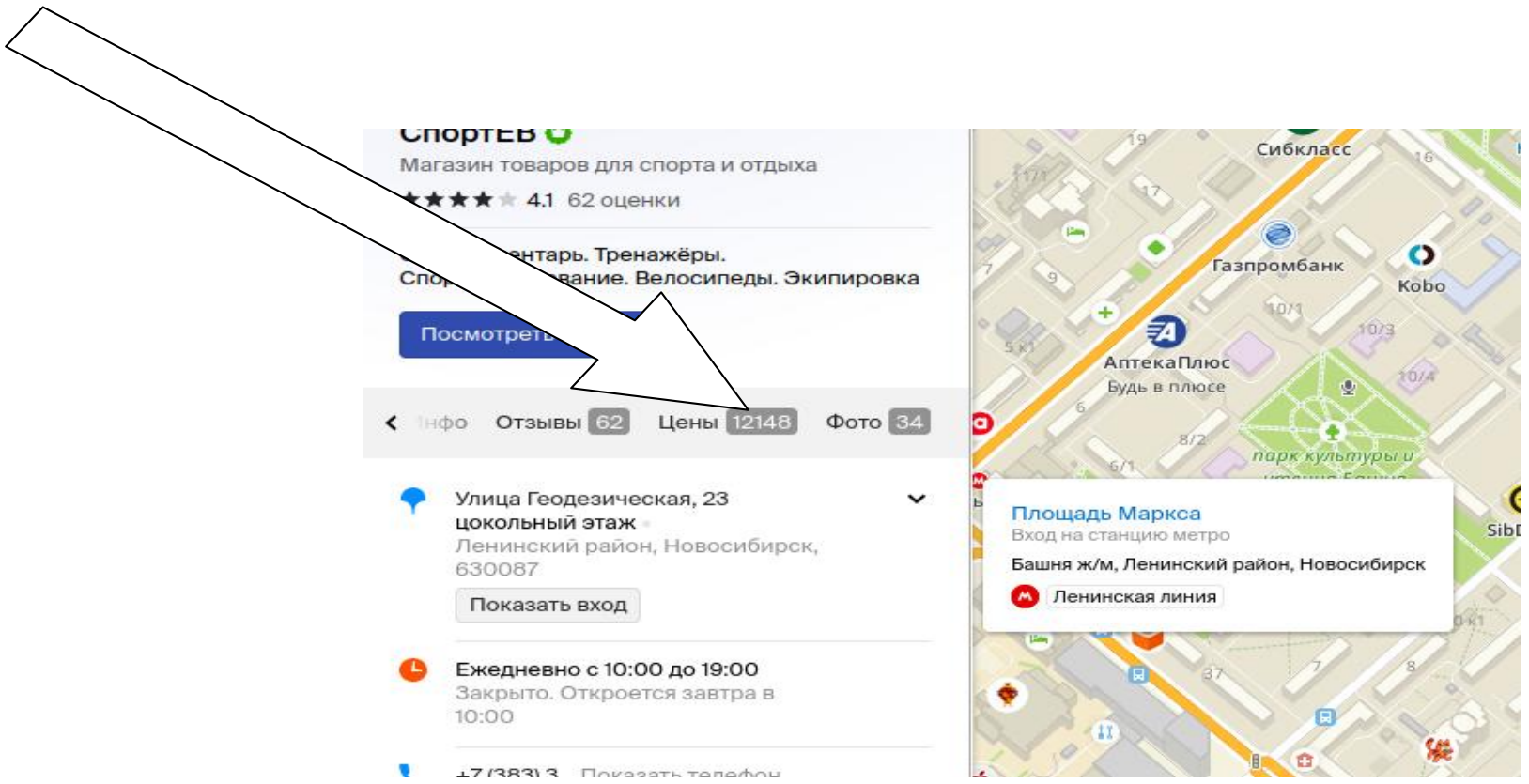
Лень, логика и работорговля
в реалиях парсинга магазинов

Олег Ануфриев, 2ГИС

RnD, NLP

БК, РК, ZX, XT… ML, AI

Вот про эту ссылочку...



Товары

- Проблема — импортируется мало
- Типовое решение — парсинг
  - Масштабирование
  - Ручной труд
- Порядка 10^6 сайтов
- Вывод — нам не подходит

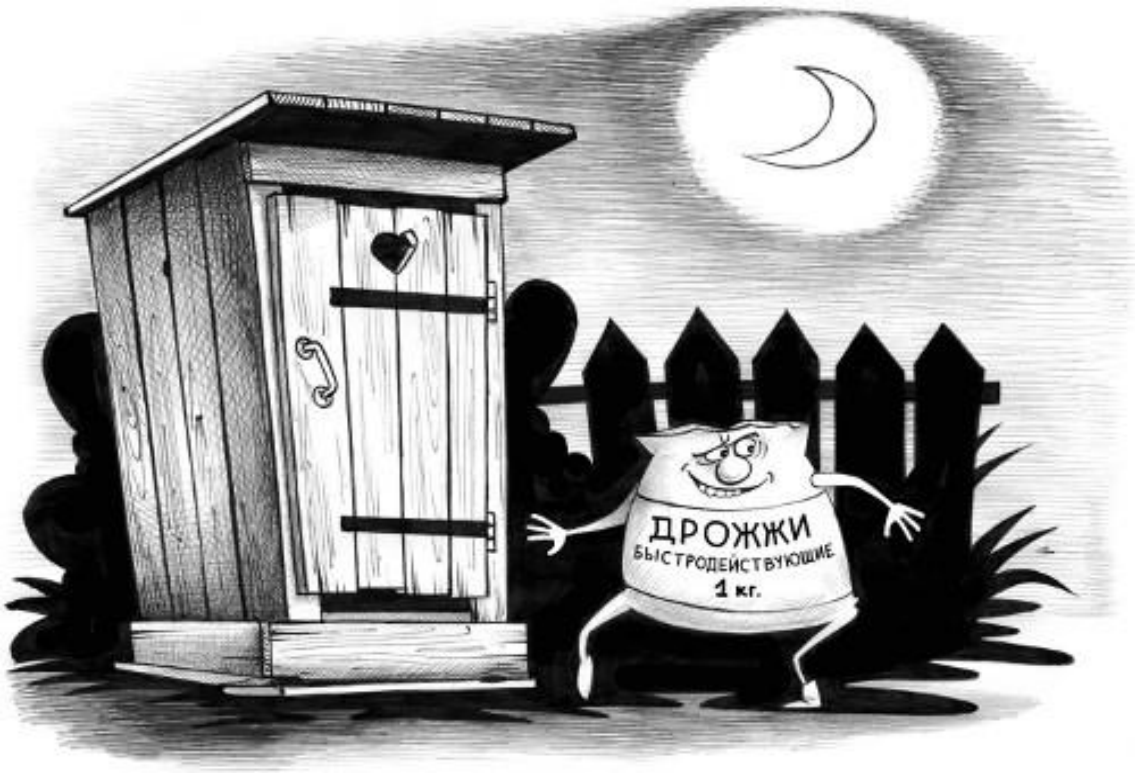
Товары и ИИ

- Машинное обучение
- Работать будет (наверное)
- Нужна разметка
- Вывод — нам не подходит
- Пока?

Хочется внести смуту

- Автоматически
- Без разметки
- Быстро

- Так не бывает!



Гипотеза 1 (Лень)

- На одном движке делают много сайтов
- Люди ленивы — магазин будет шаблонным
- Одна разметка
- Много сайтов

DOM-дерево

- Сначала html — язык разметки
- Иерархия вложенных элементов, глубокая
- Строится всё сразу и полностью
- Каждый элемент содержит часть текста
- Абзац с выделенным текстом
  - [P]text = «Абзац »
    - [SPAN] text = «с выделенным»
    - text = « текстом»

Предлагаемые решения

- XPath, lxml
- BeautifullSoup
- Scrapy

BS

```
soup = BeautifulSoup(response.text, 'lxml')
title_tag = soup.find('main').find('header').find('h1')
title_text = title_tag.text
im_src = title_tag.find('img', class_='attachment-post-image')['src']
```

да и Scrapy...

```
product_name=response.css('.item-title::text').extract()
price_range=response.css('.price-current::text').extract()

orders=response.xpath("//em[@title='Total Orders']/text()").extract()
company_name=response.xpath("//a[@class='store $p4pLog']/text()").extract()
```

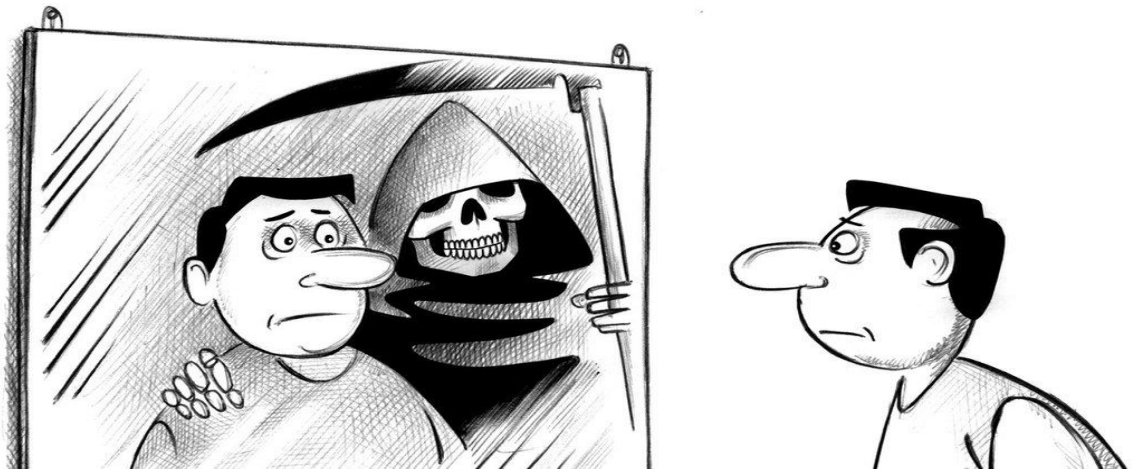
Если что, то код не наш :-)

Xpath...1.5-2%

```
self.catalog_xpaths = [
    "//*[@itemtype='http://schema.org/Store']",
    "//*[@itemtype='http://schema.org/ItemList']",
    "//div[@id='catalog']",
    "//div[@id='product-list']",
    "//div[@id='products-list']",
    "//div[contains(concat(' ', normalize-space(@class), ' '), ' catalog ')]"
]
self.catalog_item_xpaths = [
    ".//*[@itemtype='http://schema.org/Product']",
    ".//*[@itemprop='itemListElement']",
    ".//*[@class='catalog_item_content']",
]
self.catalog_item_name_xpaths = [
    ".//*[@itemprop='name']",
    ".//a[@class='item_title']",
]
self.catalog_item_offers_xpaths = [
    ".//*[@itemtype='http://schema.org/Offer']",
    ".//*[@itemprop='offers']",
    ".//*[@class='prices']",
    ".//*[@class='price']",
]
```

Xpath...1.5-2%

```
self.catalog_xpaths = [
    "//*[@itemtype='http://schema.org/Store']",
    "//*[@itemtype='http://schema.org/ItemList']",
    "//div[@id='catalog']",
    "//div[@id='product-list']",
    "//div[@id='products-list']",
    "//div[contains(concat(' ', normalize-space(@class), ' '), ' catalog ')]"
]
self.catalog_item_xpaths = [
    ".//*[@itemtype='http://schema.org/Product']",
    ".//*[@itemprop='itemListElement']",
    ".//*[@
]
```



Что вышло

- Lxml + xPath (в json)
- Поиск каталог → товар → название + цена
- Много кода для получения текста
- Что пошло не так:
  - Поражает трудолюбие кастомизаторов
  - Всё кастомизировано, а особенно — магазины!
- Вывод — 5-10% на десятках типовых шаблонов
  печальный смайлик

Плоды лени (скудные)

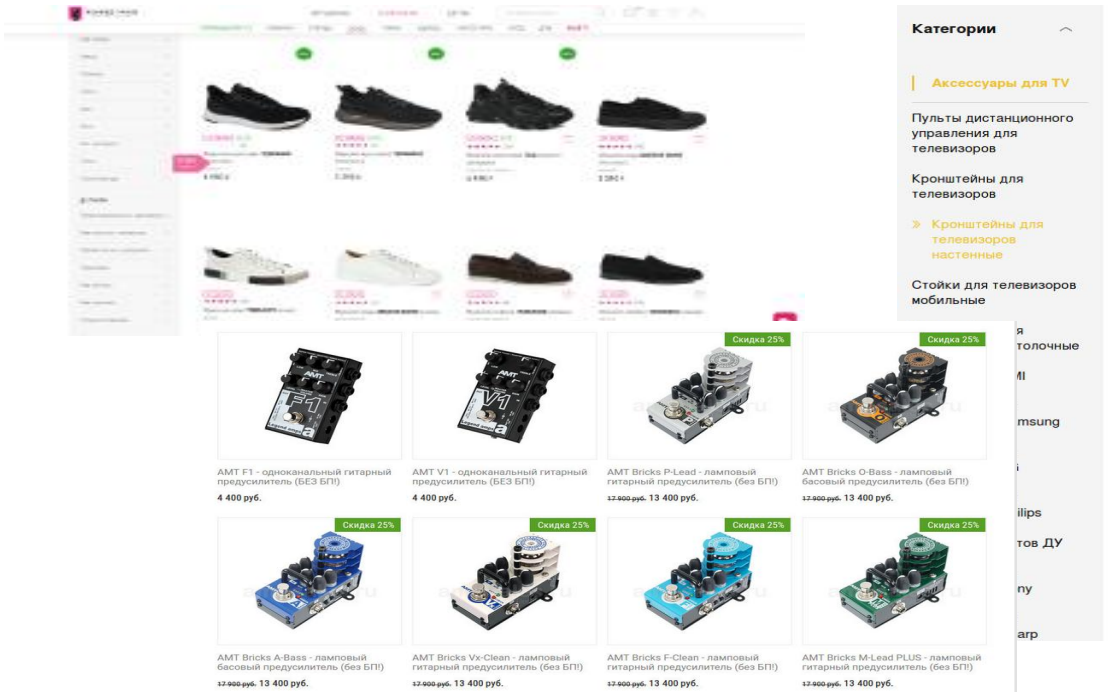
- CSS-классы несут смысл (очевидно)
- Микроразметка часто спасает
  - itemprop
  - Itemtype
  - class="price"
- Всё это — фрагментарно и вперемешку
  - class="big_red_text"
  - font color="red"
- Код работы с текстом — пригодится!

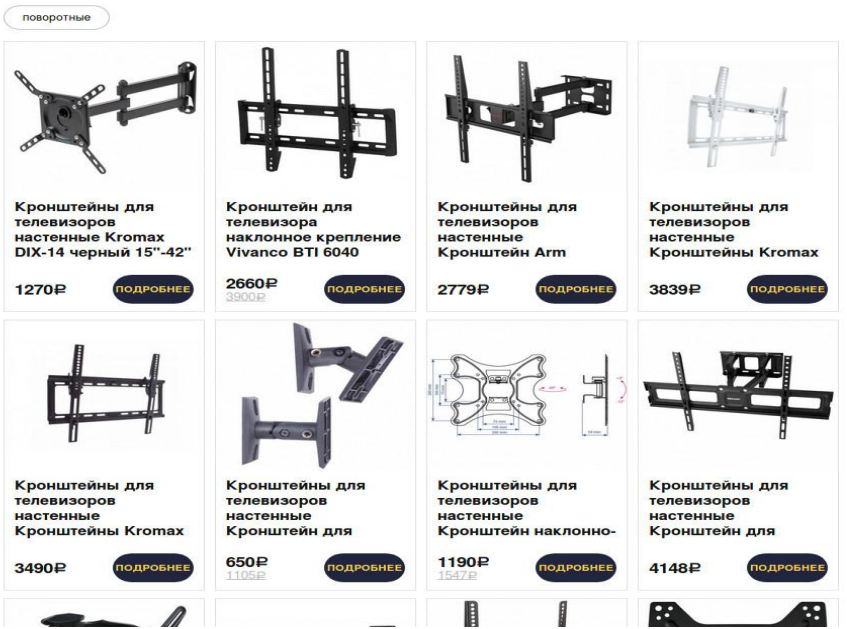
За соседним столиком

- Классификатор текстов (еда / не еда)
- Поиск цен
- Расчёт близости
- Специфика
  - Зависимость от предметной области
  - Зависимость от языка
  - Машинное обучение, разметка
- Вывод — «крепёж» не переварим

Гипотеза 2 (Логика)

- Визуально сайты логичны





Гипотеза 2 (Логика)

- Визуально и разметка логична

Гипотеза 2 (Логика)

- Визуально и разметка логична

Успешный запуск (MVP)

- Сравниваем последовательные элементы и ищем тексты с «рублями»
- Получилось!
- но МЕДЛЕННО
- и есть ошибки

Детали и проблемы

- Сравниваем по свойствам — что есть у элемента, у подэлементов и так далее (рекурсивненько)
- Нечёткое сравнение с порогом
- На всех уровнях (почти)!
- Извлекаем текст (тут - цифры и валюту)
   РУБ.
  (регулярке не разгуляться)
- Внутри могут быть невидимые и просто ненужные элементы.

Схожесть

-
  -
    -
      - <h1><a href="..">Товар 1</h1>
      - 200 руб.
    -

      ….
-
  -
    -
      - <h1><a href="..">Товар 2</h1>
      - 250 руб.

Сравнение

-
  -
    - <h1><a href="..">Товар 1</h1>
    - 200 руб.

      ….

      ….
  -
    - <h1><a href="..">Товар 2</h1>
    - 250 руб.

Извлечение текста

- <h1><a href="..">Товар 1</h1>
  - Новинка!
  - <h4 class="rating"></h4>
  - <s class="oldprice">300 руб.</s>
  - 200 руб.

    <form style="display: None" class="order_now">
    - <h3>Адрес доставки</h3>
    - <input name="shipping address">
      [sensored]
      [sensored]
      [sensored]

Мусор

- Новинка Пицца 3 Сыра [+] [-] в корзину 400 руб.
  Пицца-пирог [+] [-] в корзину 450 руб.
- Цены "по 100руб", "от 550 до 700руб.", "по 759 руб. за 1 пог. Метр"
- Скидка как цена
- Акция как часть названия.

Ошибки

- Склеились товары (табл. 2*3 или 3 товара)
  - Гайка М8 Гайка М10
  - 10руб. 12 руб. наверное со скидкой и без
- Товар разбился на два и более
  - С соусом 320руб.
  - Без соуса 300руб.
  - С двойным соусом 400руб.
- Товары не нашлись (но чаще это из-за js)
- Штраф 2000р. и другие нетовары

Три примера работорговли

- Приём специалиста
  - Пр. Переображенский — 7500руб.
  - Др. Борменталь — 3700руб.

- Вакансии
  - Слесарь КИП — 115000руб.
  - Водитель — 85000руб.

Что делать

- Оптимизация сравнений
- Оптимизация получения текста
- Фильтрация элементов
- Поиск цен++

- Не переживайте, мы из него за месяц слона сделаем!

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

Оптимизация

- <p>Корж Цена 100 <b class="p">руб.
- [p]Корж Цена 100 руб.
  [span]Цена 100 руб.
  [em]100 руб.
  [b]руб.
- tag_p has_tag_span has-1_tag_em has-2_tag_b has-2_class_p
  tag_span has_tag_em has-1_tag_b has-1_class_p
  tag_em has_tag_b has_class_p
  tag_b class_p

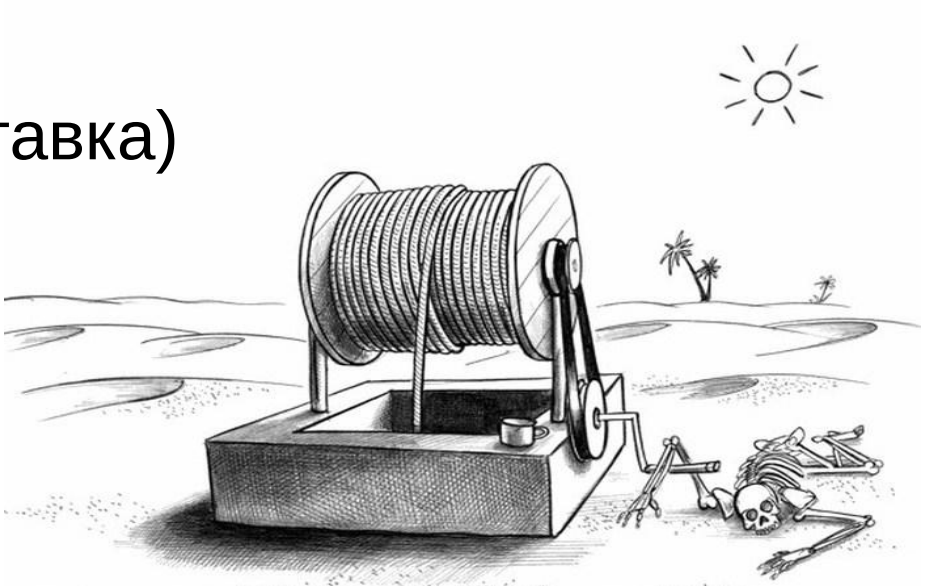
Сложное

- Чередование
  - Товар-разделитель-товар-разделитель
- Таблицы
  - Два товара в строке слиплись, строк много
  - Валюта указана в заголовке
- Копейки визуальные
  - 10000

Запускаем на AE

- Язык другой
- Другие валюты
- У верстальщиков руки под другим углом
- Иногда — цены без валюты
- Специфика — еда (кафе, доставка)

- Прокатило (в основном)!



Развитие

- Проверка → переразбиение
- Цены — 12000руб. != 17 500 руб.
- Много кода и хаков о которых я не буду рассказывать
- Много тестов с живых и уже мёртвых сайтов, которые хранят от регрессии
- Постмодерация от запрещёнки и ошибок

Нити

- Нашли похожие элементы — товары
- В товарах нашли похожие элементы
  - Цены
  - Картинки
  - Названия
  - Что-то ещё
  - Ненужное

Нити

- Корж Цена 100 руб.Купить
- Ватрушка Цена 270 руб.Купить
- Дырка бубликовая Цена 999 руб.Были, да кончилися
- Пончик Цена 150 руб.Купить

Что помогло

- Мусор — нити, бан по тексту, учёт видимости
- Склейка товаров — переразбиение
- Товары без валюты (пока выключены) — классы, нити
- Штрафы, случайные тексты — нити
- «Работорговля» - модерация

Обход дерева

- В ширину или в глубину
- Куда не ходить
- Куда ходить
- Когда остановиться

Обход дерева сайта

- Заведомо хорошие и плохие страницы

  catalog prices товары vs forum moodle контакты
- Сходство адресов, «архетипы»

  /catalog/ANY/ANY?page=ANY
- Полезные и вредные каталоги

  «Выпечка» «Гайки» vs «Новинки» «Похожие»
- Корзина резиновая, но не безразмерная
- Пагинация

Пагинация

- Сортировка
- Фильтры
- Номер страницы
  - PAGEN_1
  - PAGEN_2
  - PAGEN_3
  - PAGEN_4

Между прочим

- 20Мб страница + 500 Мб lxml + 1000 Мб «наше»
  плюс lxml «течёт»
- Свой DOM — выкинули дерево lxml — -500 МБ
- Потоковый парсер — -20 МБ
- Экономия памяти — LRU, Cython
  1000 Мб → 400

Между прочим

- Свойства - Set[str]
- Строки строятся по элементам
- Делаем унификацию — -200Мб
- Помогает LRU
- Cython (вместо Set[str])
  1000 Мб → 400

Ещё откладывается

- JS (в основном готовы)
- CSS отступычтобнеслипалось
- CSS-шрифты (для валюты)
- JS ссылки (и onclick)

- 220 тыс сайтов из 1 млн., 25 млн товаров
- Обновление раз в неделю
- 1 сек. на страницу

Лень, логика и работорговля
в реалиях парсинга магазинов

Олег Ануфриев, 2ГИС

кстати, на LightingTalk смотрите мой доклад
Цвет, это, censored, ФИЧА (AI, CV, ML)

Ожидаемые вопросы

- Языковые модели да побольше-побольше
- Что там про wget, зачем писать велосипед?
- А про индивидуальные страницы товаров?
- И про противодействие парсингу.

кстати, на LightingTalk смотрите мой доклад
Цвет, это, censored, ФИЧА (AI, CV, ML)