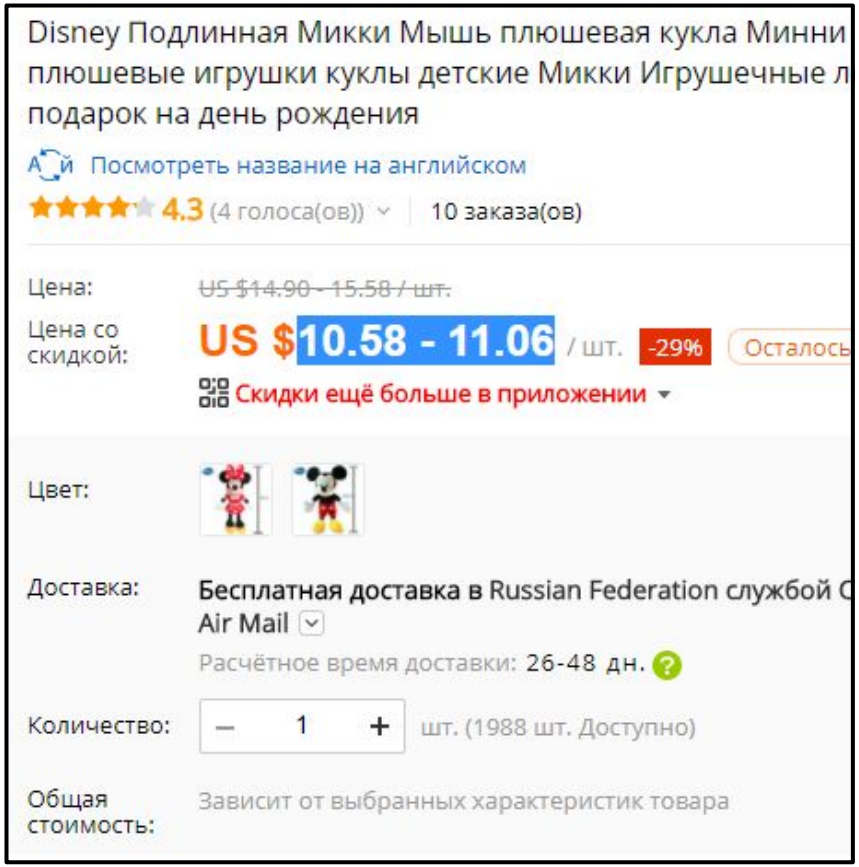
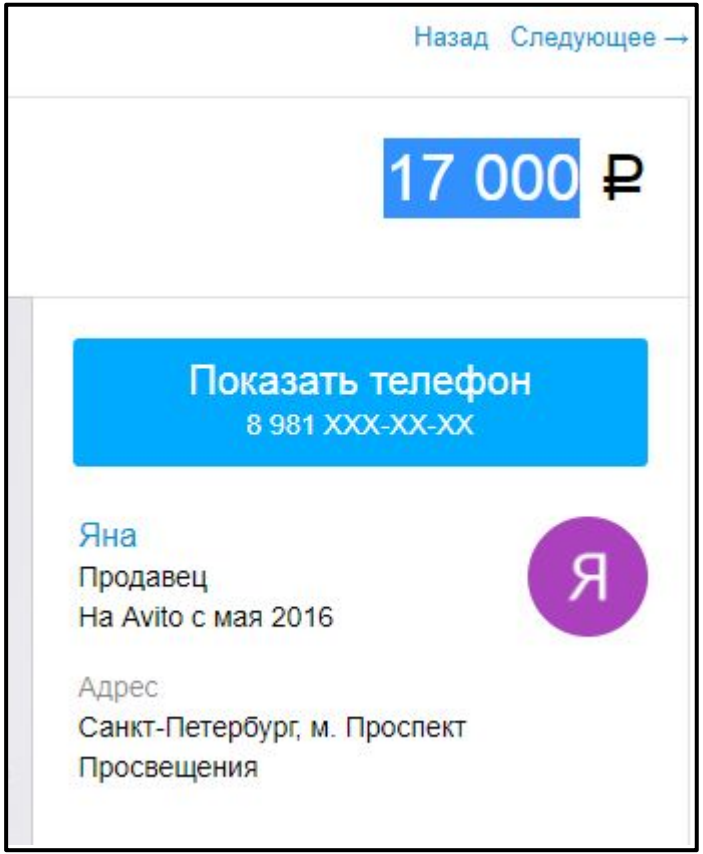
Выделение фрагментов html документа.

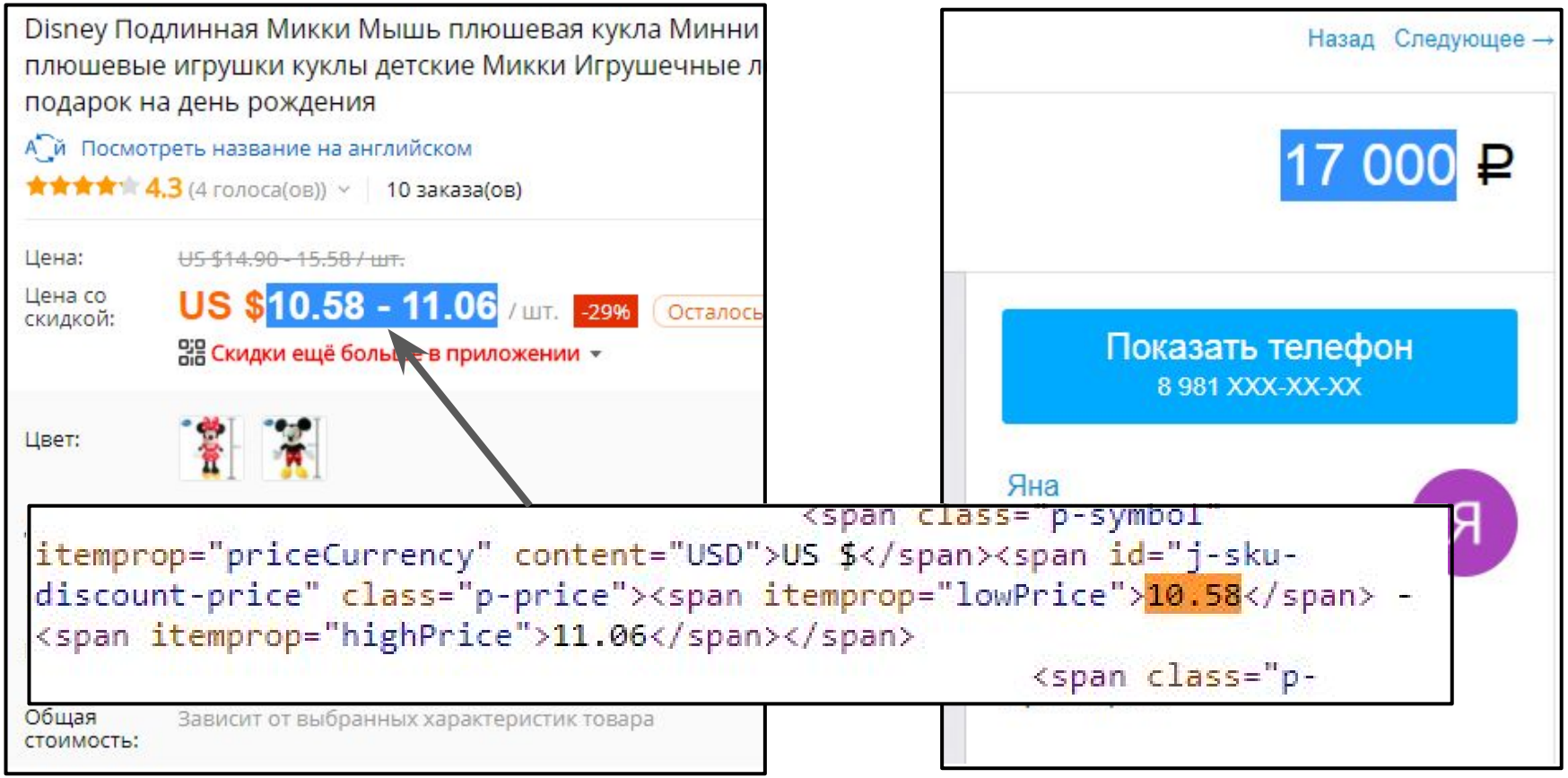
Кощенко Екатерина
Научный руководитель: Кураленок Игорь Евгеньевич

Мотивировка

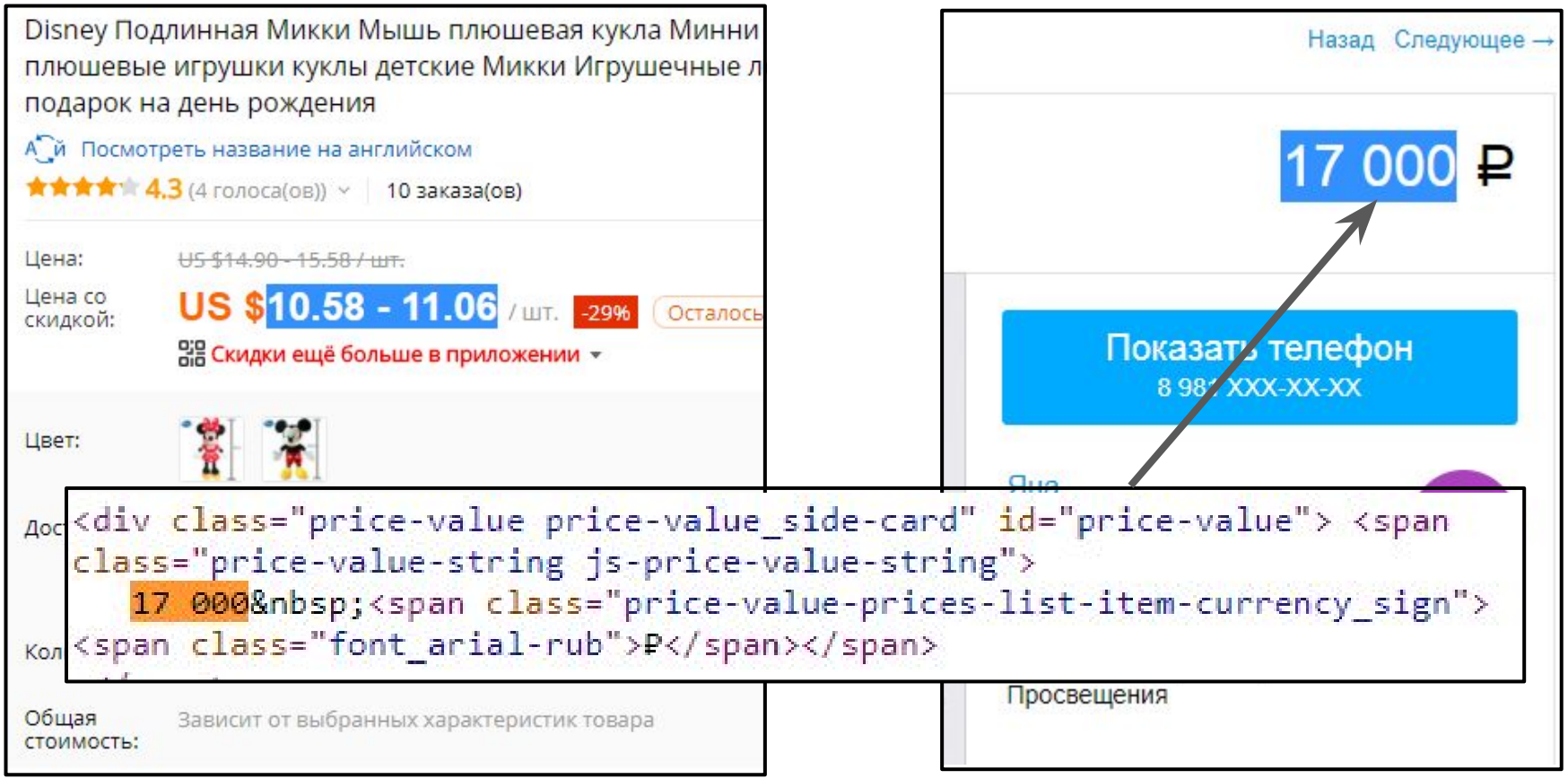




Мотивировка



Мотивировка



Задача

Постановка: дать пользователю возможность извлекать интересующую информацию из web-страниц по примеру.

Проблема: слишком различная структура html.

Методы решения в NLP

- Подобная задача в NLP — выделение частей речи.
- Старый способ решения: Hidden Markov Models.
- Более новый: Conditional Random Fields.

Решение задачи.

Решение: CRF + путь в html + классы тегов.

```
Вова пошел спать к себе домой.


<body>
        <b>Вова </b>
        <a id='verb' class='act'>
                <b>пошел </b><b>спать </b>
        </a>
        <span class='where'>к себе домой.</span>
</body>
```

```
node=Вова                path=body/b/

node=пошел               path=body/a:act/b/

node=спать               path=body/a:act/b/

node=к(/себе/домой)      path=body/span:where/
```

Conditional Random Fields

Основная формула:

$$P(y_t|y_{t-1}, x_t, coef) = \frac{\sum\limits_{val} I(y_t = val) \cdot \exp(\sum\limits_{k} \lambda_k \cdot f_k(val, (y_{t-1}), x_t))}{\sum\limits_{val} exp(\sum\limits_{k} \lambda_k \cdot f_k(val, (y_{t-1}), x_t))}$$

Что дифференцировалось:

$$FUNC = \sum_{i=1}^{Texts\_num} \sum_{t=1}^{Text\_len} \log P((x, y)_t|(x, y)_{t-1}, coef)$$

Проблема: добиться сходимости.

Результаты

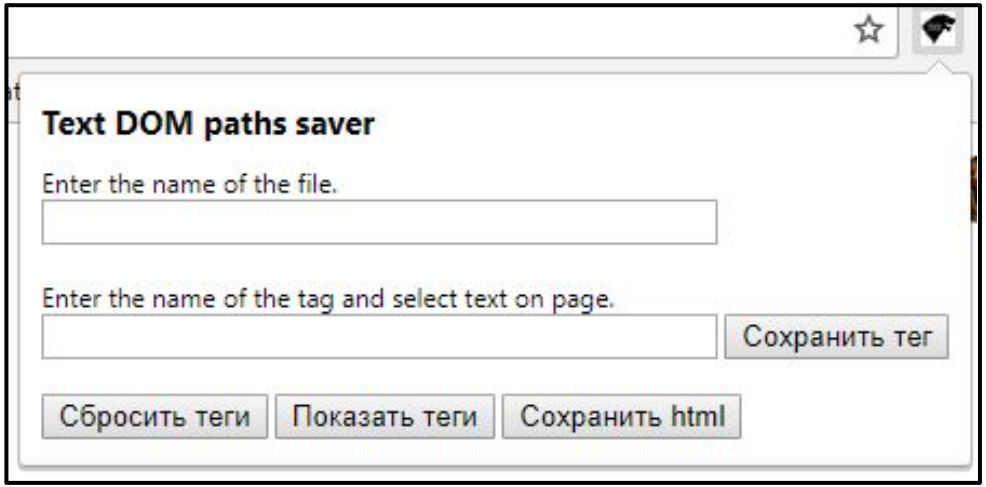
- Один сайт — успешно.
  - Сайты: aliexpress.
  - Метки: good_price, good_currency.
- Различные сайты, похожие метки — успешно.
  - Сайты: aliexpress + avito.
  - Метки: good_price, good_currency.
- Никак не связанные метки — не удалось.
  - Сайты: aliexpress + avito + formula kino.
  - Метки: good_price, good_currency, movie_name, movie_time.

Продукт

Итоговый продукт = натренированная модель CRF + Chrome плагин.

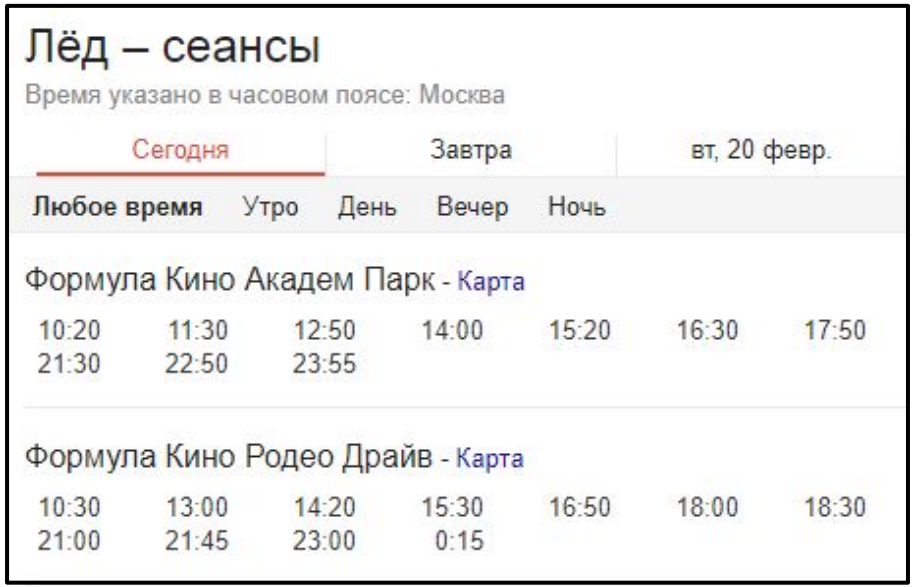
Chrome плагин:

- Присвоить выделенному на странице элементу определенную метку (train).
- Сохранить страницу в нужном для программы формате со всеми заданными метками.



Развитие

1. Улучшить реализацию CRF.
2. Развернуть сайт для удобной тренировки модели.
3. Выделение информации по шаблону по странице выдачи гугла.



Ссылки

- CRF
  - оригинальная статья
  - применение в задачах обработки текстов
  - существующая реализация
- Мой проект:
  - chrome плагин
  - реализация алгоритма CRF
  - данные

Дополнительно.

Conditional Random Fields

- Почему данная модель?
  - Наши правила распознавания будут простыми (путь, число и тд.), поэтому можно пользоваться базовым алгоритмом.
- Недостатки?
  - Вычислительная сложность анализа обучающей выборки.
  - Не работает с не встречающимися при обучении словами.
  - Не учитывается зависимость наблюдаемых элементов
- Достоинства?
  - Отсутствие проблемы смещения метки.