


# Выделение фрагментов html документа.

Кощенко Екатерина

Научный руководитель: Кураленок Игорь Евгеньевич

# Мотивировка


Disney Подлинная Микки Мышь плюшевая кукла Минни плюшевые игрушки куклы детские Микки Игрушечные л подарок на день рождения



 [Посмотреть название на английском](#)

★★★★☆ 4.3 (4 голоса(ов)) | 10 заказа(ов)


Цена: US \$14.90 - 15.58 / шт.

Цена со скидкой: **US \$10.58 - 11.06** / шт. **-29%** Осталось

 [Скидки ещё больше в приложении](#)

Цвет:  

Доставка: **Бесплатная доставка в Russian Federation службой С Air Mail**

Расчётное время доставки: 26-48 дн. 

Количество:  шт. (1988 шт. Доступно)

Общая стоимость: Зависит от выбранных характеристик товара


[Назад](#) [Следующее](#) →

**17 000** **₽**

**Показать телефон**  
8 981 XXX-XX-XX

Яна  
Продавец  
На Avito с мая 2016

Адрес  
Санкт-Петербург, м. Проспект  
Просвещения



# Мотивировка

Disney Подлинная Микки Мышь плюшевая кукла Минни плюшевые игрушки куклы детские Микки Игрушечные л подарок на день рождения

[Посмотреть название на английском](#)

★★★★☆ 4.3 (4 голоса(ов)) | 10 заказа(ов)

Цена: US \$14.90 - 15.58 / шт.

Цена со скидкой: **US \$10.58 - 11.06** / шт. -29% Осталось

Скидки ещё больше в приложении

Цвет:  

Назад Следующее →

**17 000** ₪

**Показать телефон**  
8 981 XXX-XX-XX

Яна 

```
<span class="p-symbol"
itemprop="priceCurrency" content="USD">US $</span><span id="j-sku-
discount-price" class="p-price"><span itemprop="lowPrice">10.58</span> -
<span itemprop="highPrice">11.06</span></span>
<span class="p-
```

Общая стоимость: Зависит от выбранных характеристик товара

# Мотивировка

Disney Подлинная Микки Мышь плюшевая кукла Минни плюшевые игрушки куклы детские Микки Игрушечные л подарок на день рождения

[Посмотреть название на английском](#)

★★★★☆ 4.3 (4 голоса(ов)) | 10 заказа(ов)

Цена: US \$14.90 - 15.58 / шт.

Цена со скидкой: **US \$10.58 - 11.06** / шт. -29% Осталось

Скидки ещё больше в приложении

Цвет:  

Назад Следующее →

**17 000** ₹

**Показать телефон**  
8 987 XXX-XX-XX

```
Дос <div class="price-value price-value_side-card" id="price-value"> <span  
class="price-value-string js-price-value-string">  
17 000&nbsp;&nbsp;&nbsp;<span class="price-value-prices-list-item-currency_sign">  
Кол <span class="font_arial-rub">₹</span></span>
```

Общая стоимость: Зависит от выбранных характеристик товара

Просвещения

# 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:

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

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

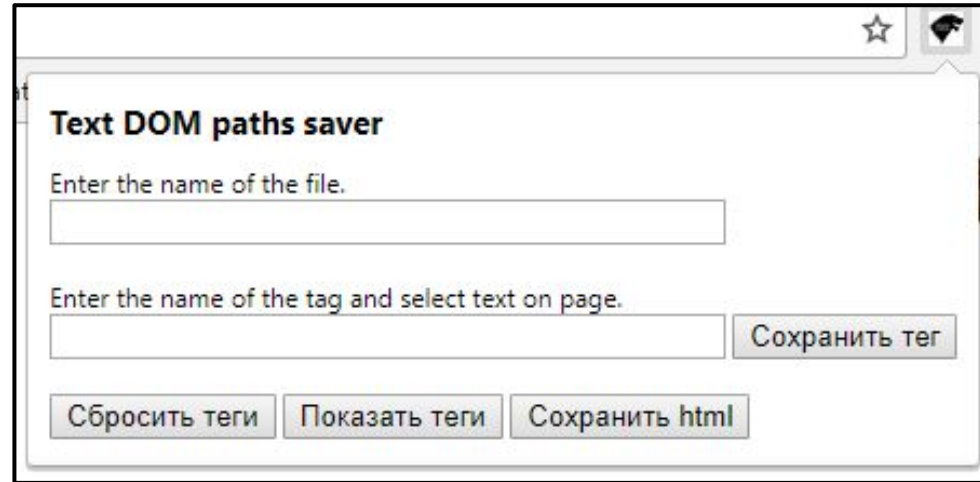
- Один сайт — успешно.
  - Сайты: aliexpress.
  - Метки: good\_price, good\_currency.
- Различные сайты, похожие метки — успешно.
  - Сайты: aliexpress + avito.
  - Метки: good\_price, good\_currency.
- Никак не связанные метки — не удалось.
  - Сайты: aliexpress + avito + formula kino.
  - Метки: good\_price, 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.

Лёд – сеансы						
Время указано в часовом поясе: Москва						
<u>Сегодня</u>		Завтра			вт, 20 февр.	
Любое время	Утро	День	Вечер	Ночь		
Формула Кино Академ Парк - <a href="#">Карта</a>						
10:20	11:30	12:50	14:00	15:20	16:30	17:50
21:30	22:50	23:55				
Формула Кино Родео Драйв - <a href="#">Карта</a>						
10:30	13:00	14:20	15:30	16:50	18:00	18:30
21:00	21:45	23:00	0:15			

# 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.