Татьяна Перевощикова
Полина Черноморченко
Полина Машковцева
БКЛ-192

Car-review corpora

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

О текстах и ходе работы

- Источник: “Поясни за тачку”
- Объем корпуса: 80126 токенов
- Первичный сбор данных: с помощью инструмента аналитики Popsters
- Обработка данных, чистка, токенизация, лемматизация с помощью Natasha, создание базы данных
- Создание веб-приложения с помощью Flask
- Реализация алгоритма поиска с помощью Python
- Встраивание алгоритма поиска в веб-приложение
- Деплой на PythonAnywhere

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Структура базы данных

1.

author_id	author
Фильтр	Фильтр
1	id226393308
2	id361554814
3	id11409697

2.

author_id	url_id	sentence_id	sentence
Фильтр	Филь...	Фильтр	Фильтр
97	1	1	Что видят люди, глядя на эво...
97	1	2	Красивый, стильный, ...
97	1	3	Что видит фанат марки, ...

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Структура базы данных

3.

url_id	url
Филь...	Фильтр
1	https://vk.com/...
2	https://vk.com/...
3	https://vk.com/...

4.

sentence_id	word	lemma	pos
Фильтр	Фильтр	Фильтр	Фил...
1	что	что	PRON
1	видят	видеть	VERB
1	люди	человек	NOUN

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

searcher.py

1. Принимаем на вход данные;
2. Определяем, сколько в запросе элементов, и разделяем их;
3. Определяем, что ищем — все словоформы, конкретную словоформу, словоформу+тег, теги…
4. Лемматизируем при необходимости;
5. Подключаемся к БД и ищем нужные нам токены;
6. Выводим результаты; сообщаем, если результатов нет

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Вопросы и решения

X: если под запрос попадают токены на стыке двух предложений, выводить оба или ни одного?
O: выводить оба.

Y: слово+POS — поиск по всем словоформам или только по данной?
O: по всем.

Z: а что с морфологической неоднозначностью?
O: у Natasha её нет :)

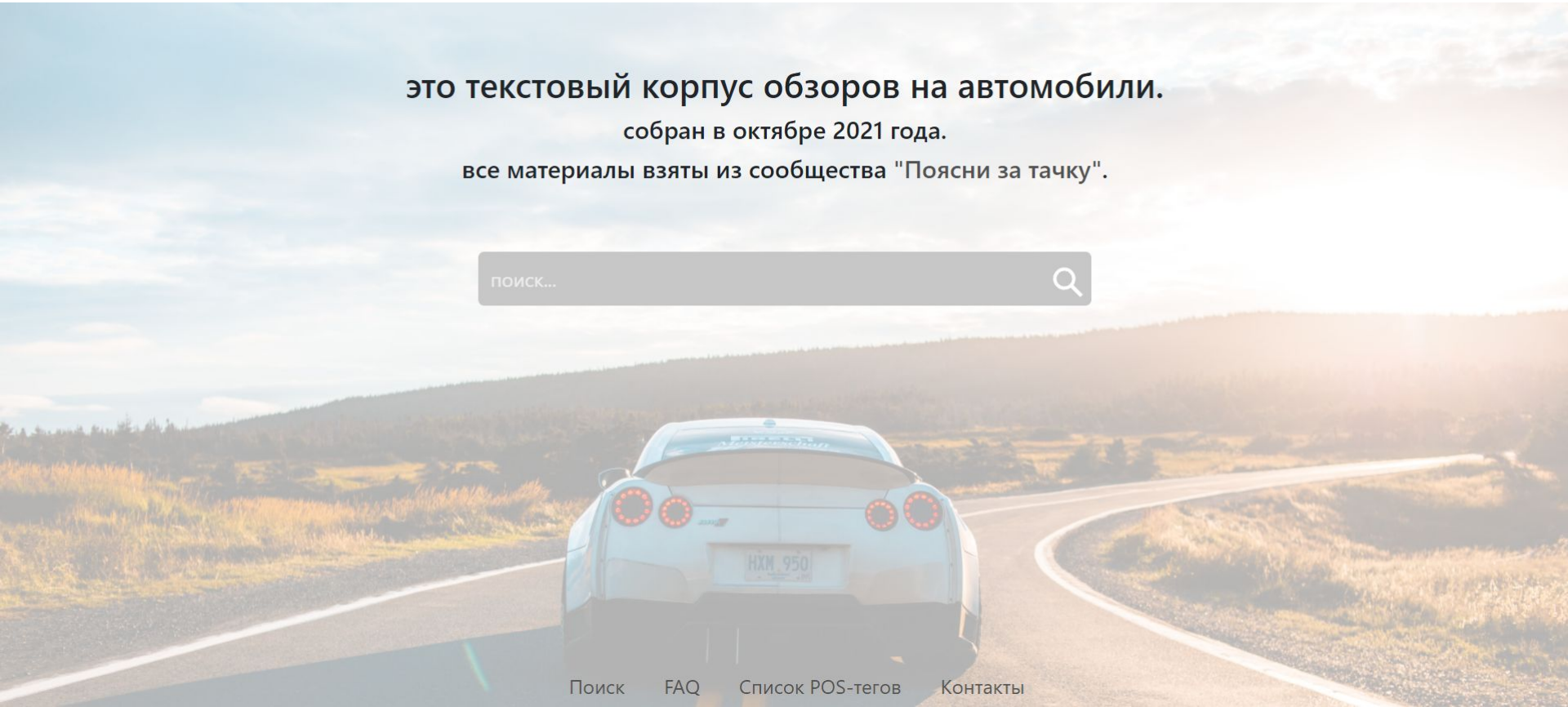
http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Примеры запросов

- данный+ADJ NOUN NOUN
- задний сиденье VERB
- колеса CCONJ
- “что” VERB машина+NOUN
- ...

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Веб-приложение



Ход работы

Таня — токенизация и лемматизация, создание БД, встраивание алгоритма поиска в веб-приложение

Полина Ч. — первичный сбор данных, реализация алгоритма поиска, деплой

Полина М. — разработка веб-приложения, презентация, деплой

Все — распределение заданий, тестирование, обсуждение ошибок и спорных случаев

http://carrevcorp.pythonanywhere.com/
https://github.com/pmashkovtseva/car-review-corpora

Спасибо за внимание!