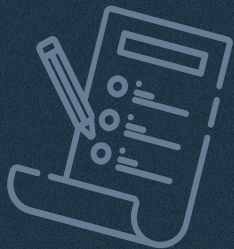


Евгений ПОНАСЕНКОВ

Первый Научный
корпус
Русского языка

<http://lerostre.pythonanywhere.com>



Сотворение проекта

1

Подумать, что вообще
я хочу сделать, как это
успеть и не умереть в
процессе



2

Сбор источников -
творчество Е. Н.
Понасенкова в
открытом доступе +
большие работы

3

Разметка и создание
БД, при помощи
rnnmorph

4

Реализация функции
поиска и интеграция в
веб-приложение
через Flask



5

Создание
нормального,
красивого, макета
сайта при помощи
bootstrap и w3
шаблонов

6

Прикручивание
всяких прикольных
мелочей типа
хайлайта, отладка,
тестирование итд



7

Отдыхать и
радоваться жизни

Материал корпуса

1. Материалы с личной страницы в
ВК

<https://vk.com/@evgenyponasenkov>

2. Личный блог Понасенкова на
mk.ru

<https://www.mk.ru/blogs/blog-evgeniya-ponasenkova.html>

3. Большие труды Понасенкова

Правда о войне 1812 года, 2004

Первая научная история войны 1812 года, 2018





Материал корпуса. Проблемы



Последней фразой, которую она пропела на публике (уже в больничной палате) была из знаменитой неаполитанской песни «Musica **proibita**» («Запретная мелодия»).

*Елена Образцова...: одним великим голосом в мире стало меньше. Блог Евгения
Понасенкова на mk.ru
12 января 2015.*

Наличие текста на других языках и прочее непотребство
латиницей





Материал корпуса. Проблемы



Номер в корпусе - 25895

От Мясницких ворот до Тверских многие здания уцелели, в том числе на бульваре дом А. А. Соловова; флигель его сгорел, также **yom** Нечаева с флигелями сгорел и Уваровой.

*Правда о войне 1812 года
2004.*



Обработано тессерактом, он в целом хорош, но некоторые косяки проскакивают



Разметка

1. На предложение побил через `razdel` и `sentenize`
2. Морфологическая разметка через `rnnmorph` и `ru morphology`, теги унифицированы, но разборы – нет

Часть речи

Граммема	Значение	Примеры
NOUN	имя существительное	хомяк
ADJ	имя прилагательное	хороший
DET	относительное местоимение	какой, который
VERB	глагол	говорю, говорит, говорил
NUM	числительное	три, пятьдесят
ADV	наречие	круто
PRON	местоимение	он
ADP	предлог	в
CONJ	союз	и
PART	частица	бы, же, лишь
INTJ	междометие	ой
PUNCT	знак препинания	., !
Н	не размечено	

Структура данных

sent_id	href	date	title	sentence
123	www.example	2013	Драйв	Как же хочется...

sent_id	lemmas	tokens	pos
123	{..., "хотеть": [1000, 1003]}	{..., "хочется": [1000]}	{..., "VERB": [1000, 1003, 1005]}

element_id	lemma	token	pos	morph
1000	хотеть	хочется	VERB	mood=Ind ...



Алгоритм поиска для одного слова



sent_df: БД с полным составом предложения - леммы, слова, части речи
word_df: БД с морфразбором каждого слова
mode: где искать: в леммах, словоформах, частях речи

```
v = np.vectorize(lambda x: element in x)
indices = np.where(v(sent_df[mode]) == True)[0]
row = sent_df.iloc[indices]
morph_indices = row[mode].apply(lambda x: x[element])
```



Алгоритм поиска для n слов

Sent_id word_id	word_1	word_2	...	word_n	Consecutive
sent_1	2	3	4	5	True
sent_2	2	8	10	1	False
....					...
sent_m	5	4	3	2	False

Пример работы поиска

Глав Поиск Запросы Топики

Введите свой запрос

е.г. ADI "молодой" + ADI NOUN Search

0.0s

Алгоритм работы поиска

```
[296]: %%time  
query = 'NOUN NOUN'  
  
results = search(query, sents, words, meta)
```

```
CPU times: total: 13.3 s  
Wall time: 13.4 s
```

```
[297]: %%time  
query = 'NOUN NOUN NOUN'  
  
results = search(query, sents, words, meta,
```

```
CPU times: total: 4min 9s  
Wall time: 4min 11s
```

```
[298]: %%time  
query = 'ADJ NOUN VERB'
```

```
results = search(query, sents, words, meta)
```

```
CPU times: total: 21.5 s  
Wall time: 21.9 s
```




Сайтик

lerostre.pythonanywhere.com

- Есть красивый титул
- Поиск работает быстро, удобно, пагинация есть, метаданные есть
- Подсказки есть
- Красивый хайлайт тоже есть, не зря же я индексы слов доставал (не работает, если вхождений $> 2000 \cdot n$)
- Морфотеги не проставлены, не влезает БД
- Запросы любой длины, но осторожно с числом результатов, если больше 400000, то упадёт



File storage: 97% full – 496.4 MB of your 512.0 MB quota [More Info](#)





Тестирование



- Длинные запросы - без разницы, если слова не слишком частые, ок
 - Загруженные запросы - типа NOUN NOUN NOUN NOUN свалят сайт, не ок
 - Лишние символы в запросе (“я!)))”) - полностью убираю всё, кроме кавычек и плюсика, ок, заглавные буквы тоже в топку
 - Большие запросы (NOUN) - NOUN имеет почти 20к вхождений, время, но только без хайлайта, иначе долго, ок
 - Не на русском языке (Napoléon) - не распарсится, слова не размечены изначально, будет пустой поиск, ок?
 - Неоднозначные запросы (знать+NOUN) - обрабатываются корректно, часть речи нормально уточняется, ок
- 
- 



Разные запросы

2018.

Номер в корпусе - 15023

«любовь к родине» — для крестьян (у которых даже не было паспортов до конца 1970-х гг. !) и пролетариев, а для детей крупных партийных функционеров, больших чинов КГБ и министерства иностранных дел, а также для послушных холопов (признаю, нередко весьма и весьма талантливых) из числа «творческой интеллигенции» — вольница за границей.

*Первая научная история войны 1812 года
2018.*

Номер в корпусе - 466

любовь к родине – для крестьян и пролетариев, а для детей больших чинов КГБ и министерства иностранных дел – вольница за границей (спросите, к примеру, депутата Алексея Митрофанова – он вам в красках расскажет).

*1812: юбилейный зомбящик, верни дворянам масленки и чулки! (26. 12. 2012). Запись в
ВК
2020-11-30.*





Разные запросы

ADJ знать+NOUN

Поиск

Найдено результатов: 9

Номер в корпусе - 22851

В Литве было учреждено Временное правительство из числа **местной знати**, которое должно было отправлять власть в тесной связи с французской оккупационной администрацией.

*Правда о войне 1812 года
2004.*

Номер в корпусе - 4906

100 Вместе с современным исследователем В. Сомовым окупемся в мир французской книги в русских собраниях: «Постепенно завоевывая русского читателя, французские книги к концу XVIII века преобладали в коллекциях императорской семьи, библиотеках **высшей знати** и столичного дворянства, причем преобладали даже над русскими.

*Первая научная история войны 1812 года
2018.*





Разные запросы

мандельштам ведь пронительнейшим образом сформулиров.

Поиск

Найдено результатов: 1

Номер в корпусе - 64

мандельштам ведь пронительнейшим образом сформулировал исторический код России: «мы живем, под собою не чуя страны».

*Кутузовский план Кремля (27. 12. 2011). Запись в ВК
2021-02-24.*



Ну вот и всё
Спасибо за внимание

