

Отчет о проекте
Сбор наиболее упоминаемых лексем
на новостном сайте lenta.ru за 2019 год

Дарья Моряшина

Задача

Собрать корпус текстов на основе новостных статей за 2019 год
Тегировать тексты на основе категорий сайта
Составить списки частотных лемм

Данные























Links_db

Источник: <https://lenta.ru/>

Date	Links
2019/01/01	['https://lenta.ru/articles/2019/01/01/analog_...
2019/01/02	['https://lenta.ru/articles/2019/01/02/nachalo...
2019/01/03	['https://lenta.ru/articles/2019/01/03/exhibit...
2019/01/04	['https://lenta.ru/articles/2019/01/04/expecta...
2019/01/05	['https://lenta.ru/photo/2019/01/05/prognoz', ...
...	...
2019/12/27	['https://lenta.ru/articles/2019/12/27/minkult...
2019/12/28	['https://lenta.ru/news/2019/12/27/gromko', 'h...
2019/12/29	['https://lenta.ru/photo/2019/12/29/frazy', 'h...
2019/12/30	['https://lenta.ru/photo/2019/12/30/grozny', '*...
2019/12/31	['https://lenta.ru/news/2019/12/30/nazvanasbor...

База текстов

Данные

Имя	Тип	Размер
 article1	Текстовый докум...	433 КБ
 article2	Текстовый докум...	456 КБ
 article3	Текстовый докум...	417 КБ
 article4	Текстовый докум...	428 КБ
 article5	Текстовый докум...	404 КБ
 article6	Текстовый докум...	402 КБ
 article7	Текстовый докум...	403 КБ
 article8	Текстовый докум...	405 КБ
 article9	Текстовый докум...	402 КБ
 article10	Текстовый докум...	403 КБ
 article11	Текстовый докум...	404 КБ
 article12	Текстовый докум...	403 КБ
 article13	Текстовый докум...	407 КБ
 article14	Текстовый докум...	404 КБ
 article15	Текстовый докум...	403 КБ
 article16	Текстовый докум...	404 КБ
 article17	Текстовый докум...	404 КБ
 article18	Текстовый докум...	405 КБ
 article19	Текстовый докум...	403 КБ
 article20	Текстовый докум...	413 КБ
 article21	Текстовый докум...	405 КБ
 article22	Текстовый докум...	413 КБ

Обработка данных: стоп слова

1. Стоп-слова из nltk
2. Топ-150 частотных лемм
“Нового частотного словаря”
О. Н. Ляшевской, С. А.
Шарова

разве, стать, надо, и, всего, хотя, либо, ради, так, около, их, я, если, него, уж, ничего, же, не, она, со, от, пожалуйста, без, чего, другой, просто, ж, за, вы, были, должен, кто, мы, но, себя, даже, другой, что, есть, без, ни, никогда,нибудь, этот, с, какая, то, прямо, для, он, сам, среди, чтоб, работа, всего, по, эти, им, хоть, от, больше, только, моя, все, этот, потом, пока, этого, над, а, даже, вы, до, три, ней, дело, сквозь, весь, ни, они, вместо, возле, во, лишь, мне, и, мой, бы, вокруг, надо, у, быть, о, опять, лучше, как, ли, тут, какой, вас, в, пусть, уже, а, так, у, время, какой, нас, ты, хотеть, ему, перед, чем, был, согласно, я, такой, они, вдруг, ладно, бы, один, два, кроме, совсем, может, всегда, уж, ним, сам, наконец, рука, как, самый, из, в, ко, об, ибо, давай, кто, потому, конечно, то, ты, этом, его, если, только, чтобы, мол, про, такой, чуть, по, еще, будто, ж, вот, нельзя, же, между, сказать, поскольку, между, во, однако, себя, знать, ведь, при, таки, когда, что, тот, раз, всех, лучше, разве, через, ничто, после, когда, сейчас, тот, ее, очень, чтобы, ну, на, до, ли, не, было, нее, них, почти, будто, еще, один, прежде, свою, тогда, день, теперь, человек, всю, можно, за, или, хоть, того, будет, где, это, через, чтоб, да, к, хорошо, зато, словно, причем, быть, под, все, зачем, при, была, ее, здесь, тем, из, раз, про, новый, идти, место, первый, свой, она, со, против, можно, мочь, два, после, благодаря, куда, который, ну, с, вам, на, да, его, более, меня, о, тебя, потом, мой, вот, том, нет, или, большой, ведь, уже, тоже, нет, говорить, себе, эту, он, жизнь, там, вроде, их, под, этой, там, над, где, именно, год, слово, чем, ей, мы, для, но, наш, впрочем, много, к, об, перед, иметь, иногда

Обработка данных: категории

```
//

    window._settings.components.customPushNotifications = {
        enabled: true,
        el: '.js-custom-push-notifications',
        title: 'Интернет и СМИ',
        slug: 'media'
    }

//]]&gt;</pre></div><div data-bbox="664 252 948 289" data-label="Text"><p>Фрагмент HTML-кода страницы</p></div><div data-bbox="78 572 607 852" data-label="Text"><pre>def category(link): # ищет название категории в коде страницы
    first = 'title: \'.+\''
    second = '[а-яА-Я ]+'
    raw = str(re.findall(first, link))
    text = re.findall(second, raw)
    result = ''
    for el in text:
        result = el.lower()
    return result</pre></div><div data-bbox="664 662 884 741" data-label="Text"><p>Поиск заголовка через<br/>регулярные выражения</p></div>
```

Обработка данных: категории

Категории на сайте:

Россия

- Мир
- Бывший СССР
- Экономика
- Силовые структуры
- Наука и техника
- Культура
- Спорт
- Интернет и СМИ
- Ценности
- Путешествия
- Из жизни
- Дом

```
army 4153
culture 4034
russia 11501
world 7436
ussr 6202
economics 6590
science 3798
sport 5688
internet 4265
values 1772
travel 2345
life 3876
home 2089
without_category 4058
```

Лемматизация и частотные списки

```
def lemma(text):  
    res = list()  
    morph = pymorphy2.MorphAnalyzer()  
    for word in text:  
        p = morph.parse(word)[0]  
        res.append(p.normal_form)  
    return res
```

```
for tag in tags:  
    with open(os.getcwd() + '\\Lemmas' + '\\{}_lemmatize.txt'.format(tag), 'r', encoding='utf8') as file:  
        corpus = file.readlines()  
        twf = defaultdict(int)  
        data = {'Word': 'Frequency'}  
        for w in corpus:  
            twf[w] += 1 / len(corpus)  
        sorted_frequency_table = sorted(twf.items(), key=itemgetter(1), reverse=True)  
        for word, freq in sorted_frequency_table[:100]:  
            appendvalue(data, word.strip('\r\n'), "%.5f" % freq)  
    save_freq(tag, data)
```

```
army 657245  
culture 602873  
russia 1499062  
world 981171  
ussr 932768  
economics 1079140  
science 545377  
sport 649030  
internet 582764  
values 243730  
travel 348087  
life 522411  
home 336332  
without_category 825589
```


Лемматизация и частотные списки

Категория: ARMY

Word Frequency

0	россия	0.00554
1	сотрудник	0.00531
2	задержать	0.00492
3	суд	0.00473
4	уголовный	0.00435
5	рубль	0.00390
6	убийство	0.00375
7	сообщить	0.00367
8	москва	0.00360
9	также	0.00355
10	данные	0.00351

Категория: RUSSIA

Word Frequency

0	россия	0.00814
1	ребёнок	0.00506
2	также	0.00469
3	заявить	0.00361
4	президент	0.00352
5	область	0.00351
6	российский	0.00327
7	ранее	0.00323
8	процент	0.00320
9	тысяча	0.00306
10	отметить	0.00289

Категория: WORLD

Word Frequency

0	сша	0.01062
1	россия	0.01060
2	страна	0.00850
3	президент	0.00705
4	российский	0.00573
5	также	0.00509
6	заявить	0.00501
7	американский	0.00496
8	военный	0.00432
9	трамп	0.00423
10	сообщать	0.00393

Категория: USSR

Word Frequency

0	украина	0.01767
1	президент	0.01100
2	украинский	0.00868
3	россия	0.00770
4	зеленский	0.00745
5	страна	0.00720
6	заявить	0.00602
7	глава	0.00524
8	киев	0.00485
9	также	0.00478
10	vladimir	0.00467

Дальнейшая работа

Сформировать корпус сленговых выражений

Сравнить корпус с собранными текстами

Написать программу, которая будет замещать сленговые выражения на стилистически нейтральные аналоги