

1.3 дубликаты

- 1 ручной поиск дубликатов;
 - 2 дубликаты с учётом регистра;
 - 3 NLTK, новая библиотека для стемминга;
 - 4 лемматизация и библиотека `rumystem3`
-

ручной поиск

1 `deduplicated().sum()`

2 `value_counts()` - наверху будут дубликаты

`.drop_duplicates()`

дубликаты и регистр

`.str.lower()`

стемминг

-нахождение основы слова

(неявные дубликаты)

vs выбирать корни самому

муха мухаммед будут неотличимы

а стемминг отличит

NLTK

1

`russian_stemmer =`

`SnowballStemmer('russian')`

2

`russian_stemmer.stem(word)`

лемматизация

-приведение слова к дефолтной форме

сущ — им падеж, единственное число;

прил — им падеж, единственное число,

мужской род;

глагол, прич, дееприч — глагол в инфинитиве

несоверш вида.

работает с несуществующими словами

```
pymystem3:  
from pymystem3 import Mystem  
m = Mystem()  
lemmas = m.lemmatize(text)
```

```
value_counts слов:  
from collections import Counter  
print(Counter(lst))
```

case: каких марок телефонов не хватает на сайте(баланс спроса и предложения)

нужно было пересчитать столбец каунт с учетом дубликатов

Нужно определить, по какой марке телефона не хватает предложений на сайте.
Помимо цифр, показывающих, как мало представлено магазинов продающих нехватящую марку,

от вас ждут доказательств, что пользователям действительно интересны малопредставленные магазинами модели.

Изучите поисковые запросы и отзывы, чтобы обосновать свои рекомендации.