1.3 дупликаты, lower, стемминг, лемматизация

1 стандартный поиск дубликатов;
2 дубликаты регистр - lower
3 NLTK -стемминг(корни)
4 лемматизация(станд.форма слова)
pymystem3

0 из практики
Когда все пропущенные/аномалии/дупликаты исправлены - нужен (повторный) чек дупликатов

Для идентификации дубликатов см совпадения по нескольким столбцам

- тем где нет missin val-ues
- в которых маловероятны совпадения!!!

---

1 стандартный поиск
1duplicated().sum()
2value_counts() - наверху будут дупликаты

.drop_duplicates()

---

2 дупликаты и регистр
.str.lower()

---

3 стемминг
-нахождение основы слова
(неявные дупликаты)

vs выбирать корни самому
муха мухаммед будут неотличимы
а стемминг отличит

NLTK
1
russian_stemmer =
SnowballStemmer('russian')
2
russian_stemmer.stem(word)

---

4 лемматизация
-приведение слова к дефолтной форме:

сущ — им падеж, единственное число;
прил — им падеж, единственное число,
мужской род;
глаг, 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 дупликатов

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

от вас ждут доказательств, что пользователям действительно интересны малопредставленные магазинами модели.
Изучите поисковые запросы и отзывы, чтобы обосновать свои рекомендации.