Поиск дубликатов

Python

Приведение строки к нижнему регистру

```
In  string.lower()
```

Подсчёт различных значений в списке

```
In  from collections import Counter

    Counter(lst)

    # используется коллекция Counter,
    # реализующая словарь для подсчёта
    # колличества неизменяемых объектов
```

Pandas

Приведение строк в колонке к нижнему регистру

```
In  data['column'].str.lower()
```

NLTK

Получение стеммера для русского языка

```
In  russian_stemmer =
                SnowballStemmer('russian')
```

Получение стема от слова на русском

```
In  russian_stemmer.stem(word)
```

PyMystem

Получение стеммера/лемматизатора для слов на русском

```
In  from pymystem3 import Mystem

    m = Mystem()
```

Лемматизация строки с русским текстом

```
In  m.lemmatize(text)
```

Словарь

Стемминг
процесс нахождения *стема* (основы слова)

Лемматизированное слово
слово, сведённое к *лемме*

Лемматизация
приведение слова к *лемме* (его словарной форме):

- *для существительных* — именительный падеж, единственное число;
- *для прилагательных* — именительный падеж, единственное число, мужской род;
- *для глаголов, причастий, деепричастий* — глагол в инфинитиве несовершенного вида.