Векторизация текстов

Получение списка лемматизированных слов

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

Получение корпуса в кодировке Юникод

In
```
corpus = data['text'].values.astype('U')
```

Поиск в тексте всех совпадений по шаблону и замена их заданной строкой

In
```
import re
re.sub(pattern,      # шаблон
       replacement,  # на что заменять
       text)         # текст, в котором искать совпадения
```

Получение мешка слов

In
```
from sklearn.feature_extraction.text import CountVectorizer

count_vect = CountVectorizer(stop_words=stopwords)  # stopwords - список стоп-слов

# bow, от англ. bag of words
bow = count_vect.fit_transform(corpus)

# словарь уникальных слов
words = count_vect.get_feature_names()
```

Получение списка N-грамм

In
```
from sklearn.feature_extraction.text import CountVectorizer

count_vect = CountVectorizer(ngram_range=(min_n, max_n))
# min_n - минимальное значение N
# max_n - максимальное значение N
```

Получение стоп-слов для русского языка

In
```
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')

stopwords = set(stopwords.words('russian'))
```

Получение TF-IDF для корпуса текста

In
```
from sklearn.feature_extraction.text import TfidfVectorizer

count_tf_idf = TfidfVectorizer(stop_words=stopwords)  # stopwords - список стоп-слов
tf_idf = count_tf_idf.fit_transform(corpus)
```

Получение TF-IDF для N-грамм корпуса текста

```
In  from sklearn.feature_extraction.text import TfidfVectorizer

    count_tf_idf = TfidfVectorizer(stop_words=stopwords, ngram_range=(min_n, max_n))
    tf_idf = count_tf_idf.fit_transform(corpus)
    # stopwords – список стоп-слов
    # min_n – минимальное значение N
    # max_n – максимальное значение N
```

Словарь

Токенизация (англ. *tokenization*)
разбиение текста на **токены**: отдельные фразы, слова, символы

Лемматизация (англ. *lemmatization*)
приведение слова к начальной форме (**лемме**)

Корпус
набор текстов, в котором эмоции и ключевые слова уже размечены

Регулярные выражения
инструмент для поиска текстов и чисел по шаблону

«Мешок слов» (англ. *bag of words*)
модель, которая преобразует текст в вектор, не учитывая порядок слов

***N*-грамма**
это последовательность из нескольких слов. *N* указывает на количество элементов и может быть любым

Анализ тональности текста, или **сентимент-анализ**, выявляет эмоционально окрашенные слова