

Обработка языка. Практика 1

Предобработка текста

Михаил Пожидаев

12 сентября 2023 г.

Токенизация

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt')
print(word_tokenize("Типа текст."))
#Ответ: ['Типа', 'текст', '.']
```

Сегментация

```
import nltk
from nltk import sent_tokenize

nltk.download('punkt')
print(sent_tokenize("Типа текст.
    Не, серьёзно."))
#Ответ: ['Типа текст.',
    'Не, серьёзно.']
```

Лемматизация

И получение грамматических атрибутов

```
import pymorphy2

m = pymorphy2.MorphAnalyzer()
w = m.parse("университетов")[0]
print(w.normal_form)
#Ответ: университет
print(w.tag)
#Ответ: NOUN, inan, masc plur, gent
```

Спасибо за внимание!

Всё о курсе: <https://marigostra.ru/materials/nlp.html>

E-mail: msp@luwrain.org

Канал в Телеграм: <https://t.me/MarigostraRu>