Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Отчёт по лабораторной работе
по дисциплине «Методы Машинного Обучения»

Выполнил:
студент группы № ИУ5-21М
Кучеренко Михаил Александрович
_____, _____

Проверил:
к.т.н., доц., Ю.Е. Гапанюк
_____, _____

2021 г.

Лабораторная работа №5

Предобработка текста.

Набор данных - Стихотворение "Простая суть" Геннадий Эсса.

Install

```
pip3 install spacy
python3 -m spacy download ru_core_news_sm
```

Задание:

Для произвольного предложения или текста решите следующие задачи:

- Токенизация.
- Частеречная разметка.
- Лемматизация.
- Выделение (распознавание) именованных сущностей.
- Разбор предложения.

```
text = '''Лишь тот способен добрым быть,
Кому пришлось от злости выть.
Лишь тот узнает дружбы цену,
Кто испытал друзей измену.
И лишь тому любовь дана,
Кто ненависть познал сполна.
Кто в жизни повстречал врагов,
Кретинов или дураков….
Где разошлись дороги наши
Из-за ненужных пустяков.
Лишь тот способен дорожить
Тем, для чего так стоит жить.
И верить, и мечтать, и ждать,
От радости своей кричать!
Искать, поверить и любить.
Для этого нам стоит быть!
Поэтому ищи, терзай,
И ни о чем не забывай.
Не жди к себе пустых признаний:
Не будет в этом оправданий.
Лишь тот способен все понять,
И строки эти прочитать.
Кому не все равно по жизни
Простую истину познать…
(С) Геннадий Эсса
'''

limit = 18
```

```
from spacy.lang.ru import Russian
import spacy
```

Токенизация

```
nlp = spacy.load('ru_core_news_sm')
spacy_text = nlp(text)
```

```
for token in spacy_text[:limit]:
    print(f'{token.text:20} - {token.pos_:7} / {token.dep_}')
```

```
Лишь                 - PART    / advmod
тот                  - DET     / nsubj
способен             - ADJ     / ROOT
добрым               - ADJ     / xcomp
быть                 - VERB    / xcomp
,                    - PUNCT   / punct

                     - SPACE   / mark
```

```
Кому                 - PRON    / iobj
пришлось             - VERB    / advcl
от                   - ADP     / case
злости               - NOUN    / obl
выть                 - VERB    / csubj
.                    - PUNCT   / punct

                   - SPACE   / advmod
Лишь                 - PART    / advmod
тот                  - DET     / nsubj
узнает               - VERB    / ROOT
дружбы               - NOUN    / iobj
```

```
spacy_text = nlp(text.replace('\n', ''))

for token in spacy_text[:limit]:
    print(f'{token.text:20} - {token.pos_:7} / {token.dep_}')
```

```
Лишь                 - PART    / advmod
тот                  - DET     / nsubj
способен             - ADJ     / ROOT
добрым               - ADJ     / xcomp
быть                 - VERB    / xcomp
,                    - PUNCT   / punct
                     - SPACE   / nsubj
Кому                 - PRON    / iobj
пришлось             - VERB    / conj
от                   - ADP     / case
злости               - NOUN    / obl
выть                 - VERB    / nmod
.                    - PUNCT   / punct
                     - SPACE   / nsubj
Лишь                 - PART    / advmod
тот                  - DET     / nsubj
узнает               - VERB    / ROOT
дружбы               - NOUN    / iobj
```

Лемматизация

```
for token in spacy_text[:limit]:
    print(token, token.lemma, token.lemma_)
```

```
Лишь 15085157332802532245 лишь
тот 9060620132269683821 тот
способен 13108567891086621765 способный
добрым 3413372487067799558 добрый
быть 13421914901540782342 быть
, 2593208677638477497 ,
  3443622981070210256
Кому 4147452868732122456 кто
пришлось 124762353642994640 58 прийтись
от 75472313111371235 81 от
злости 16727616970683336905 злость
выть 10742055882863096977 выть
. 12646065887601541794 .
  3443622981070210256
Лишь 15085157332802532245 лишь
тот 9060620132269683821 тот
узнает 98294656168915377 35 узнавать
дружбы 9180780978176652763 дружба
```

Выделение (распознавание) именованных сущностей

```
for entity in spacy_text.ents:
    print(entity.text, entity.label_)
```

```
Лишь PER
Кретинов PER
Геннадий Эсса PER
```

Разбор предложения

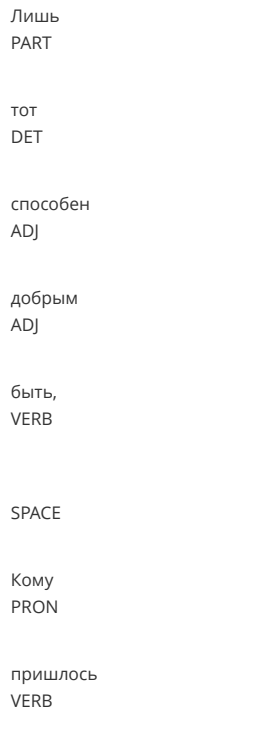
```
from spacy import displacy
```

```
sentence = len(spacy_text)
for token in spacy_text:
    if token.text == '.':
        sentence = token.i
        break
sentence += 1
```

```
displacy.render(spacy_text, style='ent', jupyter=True)
```

Лишь тот способен добрым быть, Кому пришлось от злости выть. Лишь PER тот узнает дружбы цену, Кто испытал друзей измену. И лишь тому любовь дана,Кто ненависть познал сполна.Кто в жизни повстречал врагов, Кретинов PER или дураков....Где разошлись дороги нашиИз-за ненужных пустяков.Лишь тот способен дорожитьТем, для чего так стоит жить.И верить, и мечтать, и ждать,От радости своей кричать!Искать, поверить и любить.Для этого нам стоит быть!Поэтому ищи, терзай,И ни о чем не забывай.Не жди к себе пустых признаний:Не будет в этом оправданий.Лишь тот способен все понять,И строки эти прочитать.Кому не все равно по жизниПростую истину познать...(С) Геннадий Эсса PER

```
displacy.render(spacy_text[:sentence], style='dep', jupyter=True)
```




от
ADP
злости
NOUN
выть.
VERB
advmod
nsubj
xcomp
xcomp
nsubj
iobj
conj
case
obl
nmod

```
1
```