

Министерство образования и науки 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 СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

Отчет по лабораторной работе № 5 «Предобработка текста» по курсу "Методы машинного обучения"

> Исполнитель: Студент группы ИУ5-22М Желанкина А.С. 15.05.2021

Задание лабораторной работы

Для произвольного предложения или текста решите следующие задачи:

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

Экранные формы с текстом программы и примерами её выполнения

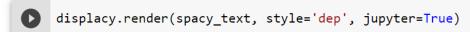
sentence = 'В 1856 году Толстой задумал повесть о возвращении декабриста из ссылки, но со временем всё больше отклонялся от первоначального замысла.'

```
токенизация
[2] from spacy.lang.ru import Russian
    import spacy
[14] nlp = spacy.load('ru_core_news_sm')
    spacy_text = nlp(sentence)
   spacy_text
    В 1856 году Толстой задумал повесть о возвращении декабриста из ссылки, но со временем всё больше отклонялся от первоначального замысла
частеречная разметка
for token in spacy_text:
       print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
    B - ADP - case
    1856 - ADJ - amod
    году - NOUN - obl
    Толстой - PROPN - nsubj
    задумал - VERB - ROOT
    повесть - NOUN - obj
    o - ADP - case
    возвращении - NOUN - nmod
    декабриста - NOUN - nmod
     из - ADP - case
    ссылки - NOUN - nmod
     , - PUNCT - punct
    HO - CCONJ - CC
    co - ADP - case
```

```
[16] for token in spacy_text:
            print(token, token.lemma, token.lemma_)
     В 15939375860797385675 в
     1856 17326095754085996097 1856
     году 10808799184780049468 год
     Толстой 4046050732273088858 толстой
     задумал 5422917566303273817 задумать
     повесть 9309007651273142624 повесть
     o 7798573245933969025 o
     возвращении 18002089832685878868 возвращение
     декабриста 4099941688933638916 декабрист
     из 12183146372738139588 из
     ссылки 5640322796510443772 ссылка
     , 2593208677638477497 ,
     но 14653780147686393572 но
     co 12039906729841018817 co
     временем 14199711609533390218 время
     всё 15417895030994739546 всё
выделение (распознавание) именованных сущностей
[17] for ent in spacy_text.ents:
        print(ent.text, ent.label_)
    Толстой PER
[18] from spacy import displacy
    displacy.render(spacy_text, style='ent', jupyter=True)
    В 1856 году Толстой рек задумал повесть о возвращении декабриста из ссылки, но со временем всё больше отклонялся от
[19] print(spacy.explain("PER"))
    Named person or family.
```

разбор предложения

[20] from spacy import displacy



 \Box

