Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа №5

«Предобработка текста»

ИСПОЛНИТЕЛЬ:

Костян Алина Алексеевна
Группа ИУ5-21М

Москва 2021

Задание:

Для произвольного предложения или текста решите следующие задачи:

1. Токенизация.
2. Частеречная разметка.
3. Лемматизация.
4. Выделение (распознавание) именованных сущностей.
5. Разбор предложения.

Текст программы:

```
In [1]: sentence = 'Мистер и миссис Дурсль проживали в доме номер четыре по Тисовой улице и всегда с гордостью заявляли,'\
        ' что они, слава богу, абсолютно нормальные люди. Уж от кого-кого, а от них никак нельзя было ожидать,'\
        ' чтобы они попали в какую-нибудь странную или загадочную ситуацию.'\
        ' Мистер и миссис Дурсль весьма неодобрительно относились к любым странностям, загадкам и прочей ерунде.'
```

токенизация

```
In [2]: from spacy.lang.ru import Russian
        import spacy
        from spacy import displacy
```

```
In [3]: nlp = spacy.load('ru_core_news_sm')
        spacy_text = nlp(sentence)
        spacy_text
```

```
Out[3]: Мистер и миссис Дурсль проживали в доме номер четыре по Тисовой улице и всегда с гордостью заявляли, что они, слава богу, абсолютно норм
        альные люди. Уж от кого-кого, а от них никак нельзя было ожидать, чтобы они попали в какую-нибудь странную или загадочную ситуацию. М
        истер и миссис Дурсль весьма неодобрительно относились к любым странностям, загадкам и прочей ерунде.
```

```
In [4]: for token in spacy_text:
            print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
Мистер - NOUN - nsubj
и - CCONJ - cc
миссис - NOUN - conj
Дурсль - PROPN - appos
проживали - VERB - ROOT
в - ADP - case
доме - NOUN - obl
номер - NOUN - nummod:entity
четыре - NUM - appos
по - ADP - case
Тисовой - ADJ - amod
улице - NOUN - nmod
и - CCONJ - cc
всегда - ADV - advmod
с - ADP - case
гордостью - NOUN - obl
заявляли - VERB - conj
, - PUNCT - punct
что - SCONJ - mark
они - PRON - nsubj
, - PUNCT - punct
слава - NOUN - parataxis
богу - NOUN - iobj
, - PUNCT - punct
абсолютно - ADV - advmod
нормальные - ADJ - amod
люди - NOUN - ccomp
. - PUNCT - punct
Уж - PART - advmod
от - ADP - case
кого - PRON - ROOT
- - PRON - nmod
кого - PRON - nmod
, - PUNCT - punct
а - CCONJ - cc
от - ADP - case
них - PRON - obl
никак - ADV - advmod
нельзя - ADV - conj
было - AUX - cop
ожидать - VERB - csubj
, - PUNCT - punct
чтобы - SCONJ - mark
они - PRON - nsubj
попали - VERB - ccomp
в - ADP - case
какую - DET - det
- - DET - det
нибудь - DET - det
странную - ADJ - amod
или - CCONJ - cc
загадочную - ADJ - conj
ситуацию - NOUN - obl
. - PUNCT - punct
чтобы - SCONJ - mark
они - PRON - nsubj
попали - VERB - ccomp
в - ADP - case
какую - DET - det
- - DET - det
нибудь - DET - det
странную - ADJ - amod
или - CCONJ - cc
загадочную - ADJ - conj
ситуацию - NOUN - obl
. - PUNCT - punct
Мистер - NOUN - nsubj
и - CCONJ - cc
миссис - NOUN - conj
Дурсль - PROPN - appos
весьма - ADV - advmod
неодобрительно - ADV - advmod
относились - VERB - ROOT
к - ADP - case
любым - DET - det
странностям - NOUN - obl
, - PUNCT - punct
загадкам - NOUN - conj
и - CCONJ - cc
прочей - ADJ - amod
ерунде - NOUN - conj
. - PUNCT - punct
```

лемматизация

In [5]:
```
for token in spacy_text:
      print(token, token.lemma, token.lemma_)
```

```
Мистер 13595807267804658451 мистер
и 15015917632809974589 и
миссис 5917637532361447116 миссис
Дурсль 14470258606626640321 дурсль
проживали 13560093007767391629 проживать
в 15939375860797385675 в
доме 17056366026889249321 дом
номер 6283387856781685377 номер
четыре 6195197003114627977 четыре
по 12047934663327436226 по
Тисовой 2788946316957815316 тисовый
улице 873128684515946723 улица
и 15015917632809974589 и
всегда 10633257961924346802 всегда
с 5863529159893111856 с
гордостью 11256727477264032983 гордость
заявляли 17644347933790284776 заявлять
, 2593208677638477497 ,
что 647928276737163203 что
они 14094787388423937788 они
, 2593208677638477497 ,
слава 29314213159180239 слава
богу 7121585943595762049 бог
, 2593208677638477497 ,
абсолютно 26493245210245994 абсолютно
нормальные 90297771621401320 80 нормальный
люди 7568775649844232870 человек
. 12646065887601541794 .
Уж 27951437888740758 84 уж
от 7547231311137123581 от
кого 4147452868732122456 кто
– 9153284864653046197 –
кого 4147452868732122456 кто
, 2593208677638477497 ,
а 17720396644590851627 а
от 7547231311137123581 от
них 17919993011152926702 них
никак 11385202641341369687 никак
нельзя 13802042952292216640 нельзя
было 6725426293294994400 было
ожидать 168416543404327949 36 ожидать
, 2593208677638477497 ,
чтобы 10327972121992521358 чтобы
они 14094787388423937788 они
попали 2830569180603168492 попасть
в 15939375860797385675 в
какую 4521607349103166500 какой
– 9153284864653046197 –
нибудь 17670556515484662119 нибудь
странную 4175539612111399975 странный
или 1530020831762146143 или
загадочную 15326335288437154292 загадочный
ситуацию 1217289187616008978 ситуация
. 12646065887601541794 .
Мистер 13595807267804658451 мистер
и 15015917632809974589 и
миссис 5917637532361447116 миссис
Дурсль 14470258606626640321 дурсль
весьма 13469450888065794792 весьма
неодобрительно 9532997101818139537 неодобрительн
относились 11984207350046093760 относиться
к 2390146911029080849 к
любым 11783268970549905777 любой
странностям 12156016637807399134 странность
, 2593208677638477497 ,
загадкам 1318552082851151575 загадка
и 15015917632809974589 и
прочей 14561572865135209795 прочей
ерунде 6955309190237157276 ерунда
. 12646065887601541794 .
```

In [6]:
```
for ent in spacy_text.ents:
    print(ent.text, ent.label_)
```

```
Дурсль PER
Тисовой улице LOC
Дурсль PER
```

In [7]:
```
displacy.render(spacy_text, style='ent', jupyter=True)
```

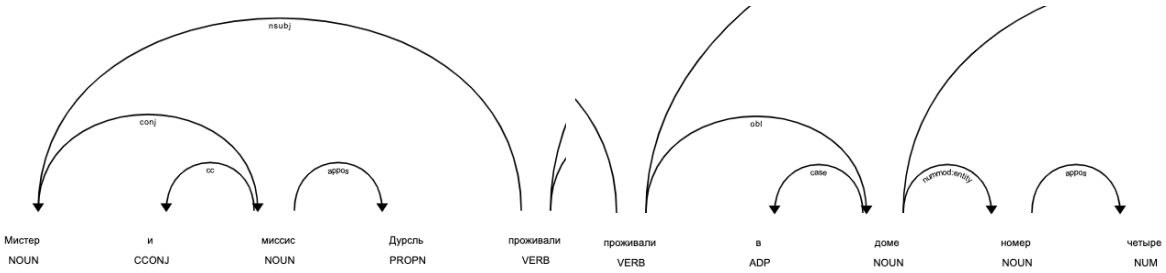
Мистер и миссис Дурсль PER проживали в доме номер четыре по Тисовой улице LOC и всегда с гордостью заявляли, что они, слава богу, абсолютно нормальные люди. Уж от кого-кого, а от них никак нельзя было ожидать, чтобы они попали в какую-нибудь странную или загадочную ситуацию. Мистер и миссис Дурсль PER весьма неодобрительно относились к любым странностям, загадкам и прочей ерунде.

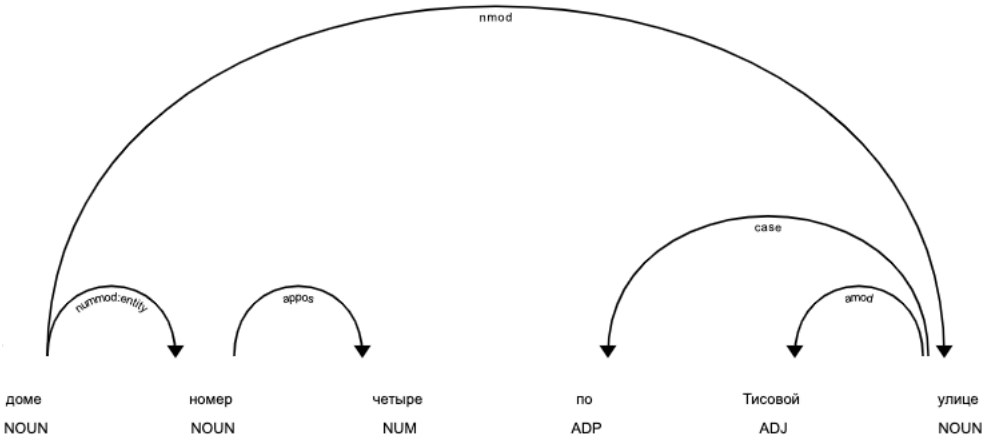
In [8]:
```
print(spacy.explain("PER"))
```

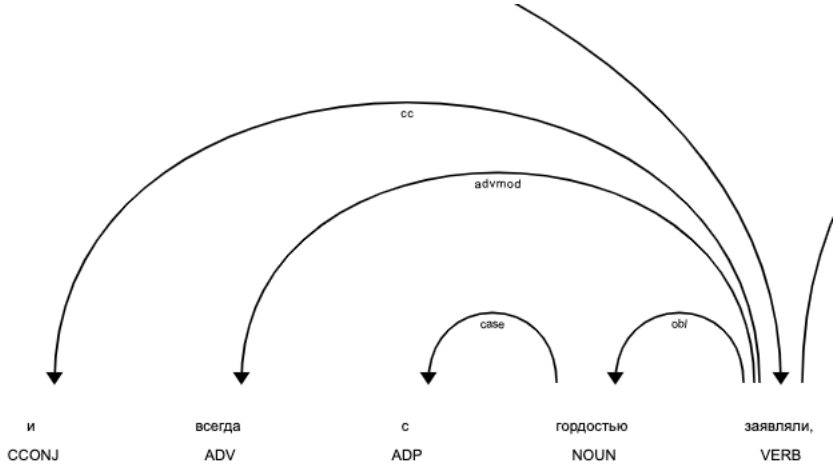
```
Named person or family.
```

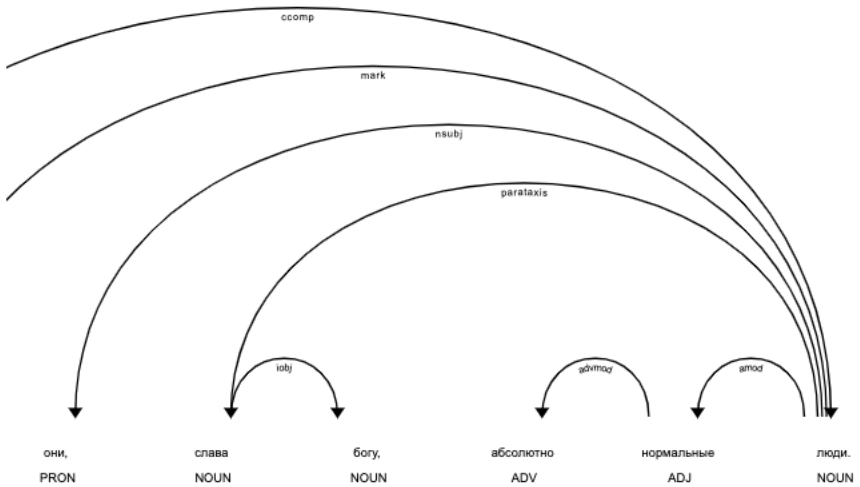
In [9]:
```
print(spacy.explain("LOC"))
```

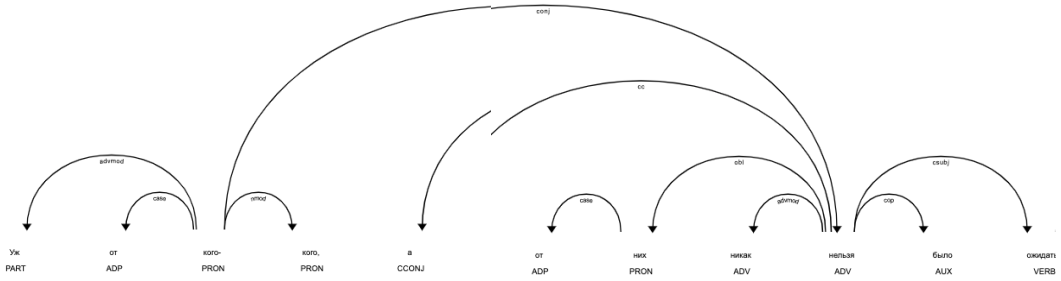
```
Non-GPE locations, mountain ranges, bodies of water
```

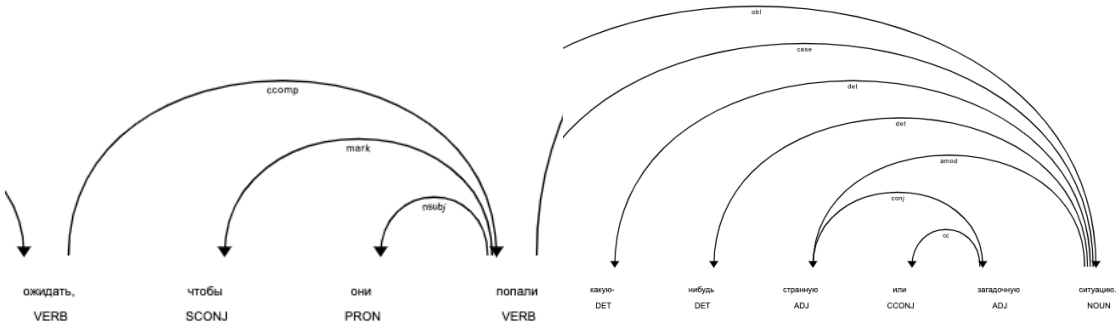

nsubj
conj
cc
appos
obl
case
nummod:entity
appos
Мистер NOUN
и CCONJ
миссис NOUN
Дурсль PROPN
проживали VERB
проживали VERB
в ADP
доме NOUN
номер NOUN
четыре NUM


nmod
nummod:entity
appos
case
amod
доме NOUN
номер NOUN
четыре NUM
по ADP
Тисовой ADJ
улице NOUN


cc
advmod
case
obl
и CCONJ
всегда ADV
с ADP
гордостью NOUN
заявляли, VERB


ccomp
mark
nsubj
parataxis
iobj
advmod
amod
они, PRON
слава NOUN
богу, NOUN
абсолютно ADV
нормальные ADJ
люди. NOUN


conj
cc
advmod
case
amod
case
obl
advmod
csubj
cop
Уж
PART
от
ADP
кого-
PRON
кого,
PRON
а
CCONJ
от
ADP
них
PRON
никак
ADV
нельзя
ADV
было
AUX
ожидать
VERB


obl
case
det
det
amod
conj
cc
ccomp
mark
nsubj
ожидать,
VERB
чтобы
SCONJ
они
PRON
попали
VERB
какую-
DET
нибудь
DET
странную
ADJ
или
CCONJ
загадочную
ADJ
ситуацию,
NOUN