Лабораторная работа №5. Предобработка текста.

Задание.

Для произвольного предложения или текста решите следующие задачи:

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

Для выполнения задания будет использваться библиотека `spacy` .

Выполнение задания.

In [1]:
```
text = "Мартин долго решал, куда пойти: в Берклейскую общедоступную читальню
```

In [2]:
```
from spacy.lang.ru import Russian
from spacy import displacy
import spacy
nlp = spacy.load('ru_core_news_sm') # Загружаем модель языка
```

Токенизация

In [3]:
```
nlp_text = nlp(text) # Возвращает Doc объект с токенезированным текстом
for token in nlp_text:
    print(token)
```

```
Мартин
долго
решал
,
куда
пойти
:
в
Берклейскую
общедоступную
читальню
или
в
Оклендскую
,
и
остановился
на
Оклендской
.
Как
знать
!
Читальня
-
```

```
самое
подходящее
для
нее
место
,
и
вполне
возможно
,
что
он
встретит
ее
там
.
```

POS. Частеречная разметка.

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

```
Мартин - PROPN - nsubj
долго - ADV - advmod
решал - VERB - ROOT
, - PUNCT - punct
куда - ADV - advmod
пойти - VERB - xcomp
: - PUNCT - punct
в - ADP - case
Берклейскую - ADJ - amod
общедоступную - ADJ - amod
читальню - NOUN - parataxis
или - CCONJ - cc
в - ADP - case
Оклендскую - ADJ - conj
, - PUNCT - punct
и - CCONJ - cc
остановился - VERB - conj
на - ADP - case
Оклендской - ADJ - obl
. - PUNCT - punct
Как - ADV - advmod
знать - VERB - ROOT
! - PUNCT - punct
Читальня - NOUN - nsubj
– - PUNCT - punct
самое - ADJ - amod
подходящее - ADJ - amod
для - ADP - case
нее - PRON - obl
место - NOUN - ROOT
, - PUNCT - punct
и - CCONJ - cc
вполне - ADV - advmod
возможно - ADJ - conj
, - PUNCT - punct
что - SCONJ - mark
он - PRON - nsubj
встретит - VERB - ccomp
ее - PRON - obj
```

```
там - ADV - advmod
. - PUNCT - punct
```

Лемматизация

In [5]:
```
for token in nlp_text:
    print("{} - {}".format(token.text, token.lemma_))
```

```
Мартин - мартин
долго - долго
решал - решать
, - ,
куда - куда
пойти - пойти
: - :
в - в
Берклейскую - берклейскую
общедоступную - общедоступный
читальню - читальня
или - или
в - в
Оклендскую - оклендский
, - ,
и - и
остановился - остановиться
на - на
Оклендской - оклендский
. - .
Как - как
знать - знать
! - !
Читальня - читальня
– - –
самое - самое
подходящее - подходящий
для - для
нее - нее
место - место
, - ,
и - и
вполне - вполне
возможно - возможный
, - ,
что - что
он - он
встретит - встретить
ее - ее
там - там
. - .
```

NER. Распознование именнованных сущностей.

In [6]:
```
displacy.render(nlp_text, style='ent', jupyter=True)
```

Мартин **PER** долго решал, куда пойти: в Берклейскую общедоступную **LOC** читальню или в Оклендскую **LOC** , и остановился на Оклендской **LOC** . Как знать!

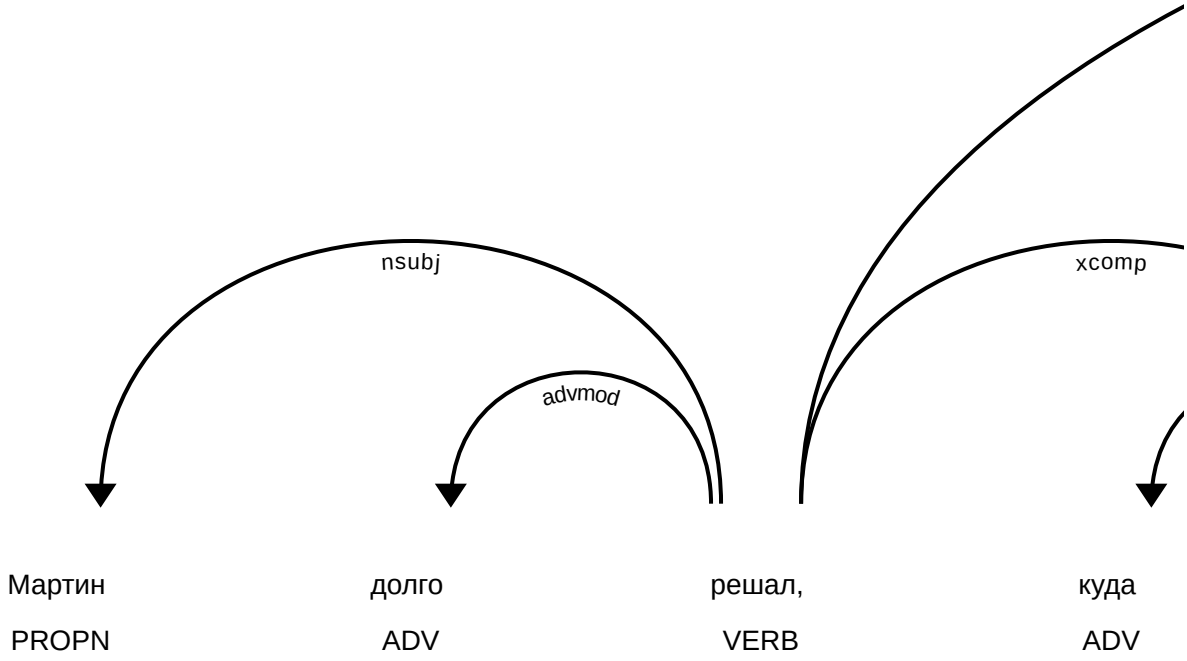
Читальня – самое подходящее для нее место, и вполне возможно, что он встретит ее там.

In [7]:
```
print(spacy.explain("LOC"))
print(spacy.explain("PER"))
```

```
Non-GPE locations, mountain ranges, bodies of water
Named person or family.
```

Разбор предложений

In [8]:
```
displacy.render(nlp_text, style='dep', jupyter=True)
```



In [9]:
```
print(spacy.explain("nsubj"))
print(spacy.explain("advmod"))
print(spacy.explain("xcomp"))
print(spacy.explain("parataxis"))
print(spacy.explain("case"))
```

```
print(spacy.explain("conj"))
print(spacy.explain("cc"))
```

```
nominal subject
adverbial modifier
open clausal complement
parataxis
case marking
conjunct
coordinating conjunction
```