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 и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

Лабораторная работа №8
По курсу
«Методы машинного обучения в АСОИУ»

«Предобработка текста»

Выполнил:

ИУ5-22М Киричков Е. Е.

27.05.2024

Проверил:

Балашов А.М.

Москва, 2024

Задание:

Для произвольного предложения или текста решите следующие задачи:

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

Токенизация

Ввод [1]:
```
text1 = 'В парке на краю города растут старые дубы и ясени. Каждый вечер туда приходи
text2 = 'В офисе всегда царит суета по утрам. Сотрудники спешат к своим рабочим места
text3 = 'В деревне, окруженной зелеными полями и лесами, все знают друг друга. Жители
```

Ввод [3]:
```
# NLTK
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/evseykirichkov/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[3]:
```
True
```

Ввод [9]:
```
from nltk import tokenize
dir(tokenize)[:20]
```

Out[9]:
```
['BlanklineTokenizer',
 'LegalitySyllableTokenizer',
 'LineTokenizer',
 'MWETokenizer',
 'NLTKWordTokenizer',
 'PunktSentenceTokenizer',
 'RegexpTokenizer',
 'ReppTokenizer',
 'SExprTokenizer',
 'SpaceTokenizer',
 'StanfordSegmenter',
 'SyllableTokenizer',
 'TabTokenizer',
 'TextTilingTokenizer',
 'ToktokTokenizer',
 'TreebankWordDetokenizer',
 'TreebankWordTokenizer',
 'TweetTokenizer',
 'WhitespaceTokenizer',
 'WordPunctTokenizer']
```

Ввод [10]:
```
nltk_tk_1 = nltk.WordPunctTokenizer()
nltk_tk_1.tokenize(text1)
```

Out[10]:
```
['В',
 'парке',
 'на',
 'краю',
 'города',
 'растут',
 'старые',
 'дубы',
 'и',
 'ясени',
 '.',
 'Каждый',
 'вечер',
 'туда',
 'приходит',
 'пожилая',
 'женщина',
 'с',
 'белой',
 'собакой',
 '.',
 'Она',
 'любит',
 'сидеть',
 'на',
 'скамейке',
 'и',
 'читать',
 'книги',
 ',',
 'пока',
 'собака',
 'бегает',
 'по',
 'лужайке',
 '.',
 'Однажды',
 'вечером',
 'к',
 'ней',
 'подошла',
 'девочка',
 'с',
 'велосипедом',
 'и',
 'спросила',
 ',',
 'не',
 'видела',
 'ли',
 'она',
 'ее',
 'пропавшего',
 'котенка',
 '.']
```

Ввод [11]:
```
# Токенизация по предложениям
nltk_tk_sents = nltk.tokenize.sent_tokenize(text1)
print(len(nltk_tk_sents))
nltk_tk_sents
```

```
4
```

Out[11]:
```
['В парке на краю города растут старые дубы и ясени.',
 'Каждый вечер туда приходит пожилая женщина с белой собакой.',
 'Она любит сидеть на скамейке и читать книги, пока собака бегает по лужайке.',
 'Однажды вечером к ней подошла девочка с велосипедом и спросила, не видела ли она е
е пропавшего котенка.']
```

Ввод [15]:
```
# Spacy
# Установка библиотеки spacy
! pip install spacy
# Установка русской модели
! python -m spacy download ru_core_news_sm
```

```
Requirement already satisfied: spacy in /Users/evseykirichkov/anaconda3/lib/pytho
n3.11/site-packages (3.7.4)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /Users/evseykirichk
ov/anaconda3/lib/python3.11/site-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /Users/evseykirichk
ov/anaconda3/lib/python3.11/site-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/evseykirichko
v/anaconda3/lib/python3.11/site-packages (from spacy) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /Users/evseykirichkov/anaco
nda3/lib/python3.11/site-packages (from spacy) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/evseykirichkov/ana
conda3/lib/python3.11/site-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.3.0,>=8.2.2 in /Users/evseykirichkov/anaco
nda3/lib/python3.11/site-packages (from spacy) (8.2.3)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /Users/evseykirichkov/anac
onda3/lib/python3.11/site-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /Users/evseykirichkov/anaco
nda3/lib/python3.11/site-packages (from spacy) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /Users/evseykirichkov/a
```

Ввод [16]:
```
from spacy.lang.ru import Russian
import spacy
nlp = spacy.load('ru_core_news_sm')
spacy_text1 = nlp(text1)
spacy_text1
```

Out[16]:
```
В парке на краю города растут старые дубы и ясени. Каждый вечер туда приходит пожила
я женщина с белой собакой. Она любит сидеть на скамейке и читать книги, пока собака
бегает по лужайке. Однажды вечером к ней подошла девочка с велосипедом и спросила, н
е видела ли она ее пропавшего котенка.
```

Ввод [17]:
```
for t in spacy_text1:
    print(t)
```

```
В
парке
на
краю
города
растут
старые
дубы
и
ясени
.
Каждый
вечер
туда
приходит
пожилая
женщина
с
белой
собакой
.
Она
любит
сидеть
на
скамейке
и
читать
книги
,
пока
собака
бегает
по
лужайке
.
Однажды
вечером
к
ней
подошла
девочка
с
велосипедом
и
спросила
,
не
видела
ли
она
ее
пропавшего
котенка
.
```

Ввод [18]:
```
spacy_text2 = nlp(text2)
spacy_text2
```

Out[18]:
```
В офисе всегда царит суета по утрам. Сотрудники спешат к своим рабочим местам, на ст
олах дымятся чашки с горячим кофе, и слышатся звуки печатающих клавиатур. Мария, нов
ая сотрудница, нервно листала страницы своего блокнота, готовясь к первой встрече с
клиентами. Ее коллеги подбадривали ее, уверяя, что все пройдет отлично.
```

Ввод [19]:
```
spacy_text3 = nlp(text3)
spacy_text3
```

Out[19]:
```
В деревне, окруженной зелеными полями и лесами, все знают друг друга. Жители каждое
воскресенье собираются на площади у старой церкви, чтобы обсудить новости и поделить
ся планами. В это воскресенье все обсуждали предстоящий фестиваль урожая. Дети радов
ались, предвкушая конкурсы и угощения, а взрослые готовились к выставке лучших сельс
кохозяйственных продуктов.
```

Ввод [22]:
```
# Natasha
! pip install razdel
```

```
Collecting razdel
  Obtaining dependency information for razdel from https://files.pythonhosted.org/pa
ckages/15/2c/664223a3924aa6e70479f7d37220b3a658765b9cfe760b4af7ffdc50d38f/razdel-0.
5.0-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/15/2c/664223a
3924aa6e70479f7d37220b3a658765b9cfe760b4af7ffdc50d38f/razdel-0.5.0-py3-none-any.whl.
metadata)
  Downloading razdel-0.5.0-py3-none-any.whl.metadata (10.0 kB)
Downloading razdel-0.5.0-py3-none-any.whl (21 kB)
Installing collected packages: razdel
Successfully installed razdel-0.5.0
```

Ввод [23]:
```
from razdel import tokenize, sentenize
```

Ввод [24]:
```
n_tok_text1 = list(tokenize(text1))
n_tok_text1
```

Out[24]:
```
[Substring(0, 1, 'В'),
 Substring(2, 7, 'парке'),
 Substring(8, 10, 'на'),
 Substring(11, 15, 'краю'),
 Substring(16, 22, 'города'),
 Substring(23, 29, 'растут'),
 Substring(30, 36, 'старые'),
 Substring(37, 41, 'дубы'),
 Substring(42, 43, 'и'),
 Substring(44, 49, 'ясени'),
 Substring(49, 50, '.'),
 Substring(51, 57, 'Каждый'),
 Substring(58, 63, 'вечер'),
 Substring(64, 68, 'туда'),
 Substring(69, 77, 'приходит'),
 Substring(78, 85, 'пожилая'),
 Substring(86, 93, 'женщина'),
 Substring(94, 95, 'с'),
 Substring(96, 101, 'белой'),
 Substring(102, 109, 'собакой'),
 Substring(109, 110, '.'),
 Substring(111, 114, 'Она'),
 Substring(115, 120, 'любит'),
 Substring(121, 127, 'сидеть'),
 Substring(128, 130, 'на'),
 Substring(131, 139, 'скамейке'),
 Substring(140, 141, 'и'),
 Substring(142, 148, 'читать'),
 Substring(149, 154, 'книги'),
 Substring(154, 155, ','),
 Substring(156, 160, 'пока'),
 Substring(161, 167, 'собака'),
 Substring(168, 174, 'бегает'),
 Substring(175, 177, 'по'),
 Substring(178, 185, 'лужайке'),
 Substring(185, 186, '.'),
 Substring(187, 194, 'Однажды'),
 Substring(195, 202, 'вечером'),
 Substring(203, 204, 'к'),
 Substring(205, 208, 'ней'),
 Substring(209, 216, 'подошла'),
 Substring(217, 224, 'девочка'),
 Substring(225, 226, 'с'),
 Substring(227, 238, 'велосипедом'),
 Substring(239, 240, 'и'),
 Substring(241, 249, 'спросила'),
 Substring(249, 250, ','),
 Substring(251, 253, 'не'),
 Substring(254, 260, 'видела'),
 Substring(261, 263, 'ли'),
 Substring(264, 267, 'она'),
 Substring(268, 270, 'ее'),
 Substring(271, 281, 'пропавшего'),
 Substring(282, 289, 'котенка'),
 Substring(289, 290, '.')]
```

Ввод [25]:
```
[_.text for _ in n_tok_text1]
```

Out[25]:
```
['В',
 'парке',
 'на',
 'краю',
 'города',
 'растут',
 'старые',
 'дубы',
 'и',
 'ясени',
 '.',
 'Каждый',
 'вечер',
 'туда',
 'приходит',
 'пожилая',
 'женщина',
 'с',
 'белой',
 'собакой',
 '.',
 'Она',
 'любит',
 'сидеть',
 'на',
 'скамейке',
 'и',
 'читать',
 'книги',
 ',',
 'пока',
 'собака',
 'бегает',
 'по',
 'лужайке',
 '.',
 'Однажды',
 'вечером',
 'к',
 'ней',
 'подошла',
 'девочка',
 'с',
 'велосипедом',
 'и',
 'спросила',
 ',',
 'не',
 'видела',
 'ли',
 'она',
 'ее',
 'пропавшего',
 'котенка',
 '.']
```

Ввод [26]:
```
n_sen_text1 = list(sentenize(text1))
n_sen_text1
```

Out[26]:
```
[Substring(0, 50, 'В парке на краю города растут старые дубы и ясени.'),
 Substring(51,
           110,
           'Каждый вечер туда приходит пожилая женщина с белой собакой.'),
 Substring(111,
           186,
           'Она любит сидеть на скамейке и читать книги, пока собака бегает по лужай
ке.'),
 Substring(187,
           290,
           'Однажды вечером к ней подошла девочка с велосипедом и спросила, не видел
а ли она ее пропавшего котенка.')]
```

Ввод [27]:
```
[_.text for _ in n_sen_text1], len([_.text for _ in n_sen_text1])
```

Out[27]:
```
(['В парке на краю города растут старые дубы и ясени.',
  'Каждый вечер туда приходит пожилая женщина с белой собакой.',
  'Она любит сидеть на скамейке и читать книги, пока собака бегает по лужайке.',
  'Однажды вечером к ней подошла девочка с велосипедом и спросила, не видела ли она
ее пропавшего котенка.'],
 4)
```

Ввод [28]:
```
# Этот вариант токенизации нужен для последующей обработки
def n_sentenize(text):
    n_sen_chunk = []
    for sent in sentenize(text):
        tokens = [_.text for _ in tokenize(sent.text)]
        n_sen_chunk.append(tokens)
    return n_sen_chunk
```

Ввод [29]:
```
n_sen_chunk_1 = n_sentenize(text1)
n_sen_chunk_1
```

Out[29]:
```
[['В',
  'парке',
  'на',
  'краю',
  'города',
  'растут',
  'старые',
  'дубы',
  'и',
  'ясени',
  '.'],
 ['Каждый',
  'вечер',
  'туда',
  'приходит',
  'пожилая',
  'женщина',
  'с',
  'белой',
  'собакой',
  '.'],
 ['Она',
  'любит',
  'сидеть',
  'на',
  'скамейке',
  'и',
  'читать',
  'книги',
  ',',
  'пока',
  'собака',
  'бегает',
  'по',
  'лужайке',
  '.'],
 ['Однажды',
  'вечером',
  'к',
  'ней',
  'подошла',
  'девочка',
  'с',
  'велосипедом',
  'и',
  'спросила',
  ',',
  'не',
  'видела',
  'ли',
  'она',
  'ее',
  'пропавшего',
  'котенка',
  '.']]
```

Частеречная разметка

Ввод [33]:
```
# Spacy
for token in spacy_text1:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
В - ADP - case
парке - NOUN - obl
на - ADP - case
краю - NOUN - nmod
города - NOUN - nmod
растут - VERB - ROOT
старые - ADJ - amod
дубы - NOUN - nsubj
и - CCONJ - cc
ясени - NOUN - conj
. - PUNCT - punct
Каждый - DET - det
вечер - NOUN - obl
туда - ADV - advmod
приходит - VERB - ROOT
пожилая - ADJ - amod
женщина - NOUN - nsubj
с - ADP - case
белой - ADJ - amod
собакой - NOUN - nmod
. - PUNCT - punct
Она - PRON - nsubj
любит - VERB - ROOT
сидеть - VERB - xcomp
на - ADP - case
скамейке - NOUN - obl
и - CCONJ - cc
читать - VERB - conj
книги - NOUN - obj
, - PUNCT - punct
пока - SCONJ - mark
собака - NOUN - nsubj
бегает - VERB - advcl
по - ADP - case
лужайке - NOUN - obl
. - PUNCT - punct
Однажды - ADV - advmod
вечером - NOUN - obl
к - ADP - case
ней - PRON - obl
подошла - VERB - ROOT
девочка - NOUN - nsubj
с - ADP - case
велосипедом - NOUN - nmod
и - CCONJ - cc
спросила - VERB - conj
, - PUNCT - punct
не - PART - advmod
видела - VERB - ccomp
ли - PART - advmod
она - PRON - nsubj
ее - DET - det
пропавшего - VERB - amod
котенка - NOUN - obj
. - PUNCT - punct
```

Ввод [37]:
```
# Natasha
! pip install navec
! pip install slovnet
```

```
Requirement already satisfied: navec in /Users/evseykirichkov/anaconda3/lib/python3.
11/site-packages (0.10.0)
Requirement already satisfied: numpy in /Users/evseykirichkov/anaconda3/lib/python3.
11/site-packages (from navec) (1.26.4)
Collecting slovnet
  Obtaining dependency information for slovnet from https://files.pythonhosted.org/p
ackages/7c/32/d5aff64e3d51ec4021674215680f16b7d2907860c6443b0d058579ac7d59/slovnet-
0.6.0-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/7c/32/d5aff
64e3d51ec4021674215680f16b7d2907860c6443b0d058579ac7d59/slovnet-0.6.0-py3-none-any.w
hl.metadata)
  Downloading slovnet-0.6.0-py3-none-any.whl.metadata (34 kB)
Requirement already satisfied: numpy in /Users/evseykirichkov/anaconda3/lib/python3.
11/site-packages (from slovnet) (1.26.4)
Requirement already satisfied: razdel in /Users/evseykirichkov/anaconda3/lib/python
3.11/site-packages (from slovnet) (0.5.0)
Requirement already satisfied: navec in /Users/evseykirichkov/anaconda3/lib/python3.
11/site-packages (from slovnet) (0.10.0)
Downloading slovnet-0.6.0-py3-none-any.whl (46 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.7/46.7 kB 1.4 MB/s eta 0:00:00
Installing collected packages: slovnet
Successfully installed slovnet-0.6.0
```

Ввод [38]:
```
from navec import Navec
from slovnet import Morph
```

Ввод [39]:
```
# Файл необходимо скачать по ссылке https://github.com/natasha/navec#downloads
navec = Navec.load('data/navec_news_v1_1B_250K_300d_100q.tar')
```

Ввод [40]:
```
# Файл необходимо скачать по ссылке https://github.com/natasha/slovnet#downloads
n_morph = Morph.load('data/slovnet_morph_news_v1.tar', batch_size=4)
```

Ввод [41]:
```
morph_res = n_morph.navec(navec)
```

Ввод [42]:
```
def print_pos(markup):
    for token in markup.tokens:
        print('{} - {}'.format(token.text, token.tag))
```

Ввод [43]:
```
n_text1_markup = list(_ for _ in n_morph.map(n_sen_chunk_1))
[print_pos(x) for x in n_text1_markup]
```

```
В – ADP
парке – NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
на – ADP
краю – NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
города – NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
растут – VERB|Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice
=Act
старые – ADJ|Case=Nom|Degree=Pos|Number=Plur
дубы – NOUN|Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur
и – CCONJ
ясени – NOUN|Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur
. – PUNCT
Каждый – DET|Case=Acc|Gender=Masc|Number=Sing
вечер – NOUN|Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing
туда – ADV|Degree=Pos
приходит – VERB|Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voi
ce=Act
пожилая – ADJ|Case=Nom|Degree=Pos|Gender=Fem|Number=Sing
женщина – NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing
с – ADP
белой – ADJ|Case=Ins|Degree=Pos|Gender=Fem|Number=Sing
собакой – NOUN|Animacy=Anim|Case=Ins|Gender=Fem|Number=Sing
. – PUNCT
Она – PRON|Case=Nom|Gender=Fem|Number=Sing|Person=3
любит – VERB|Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=
Act
сидеть – VERB|Aspect=Imp|VerbForm=Inf|Voice=Act
на – ADP
скамейке – NOUN|Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing
и – CCONJ
читать – VERB|Aspect=Imp|VerbForm=Inf|Voice=Act
книги – NOUN|Animacy=Inan|Case=Acc|Gender=Fem|Number=Plur
, – PUNCT
пока – SCONJ
собака – NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing
бегает – VERB|Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice
=Act
по – ADP
лужайке – NOUN|Animacy=Inan|Case=Dat|Gender=Fem|Number=Sing
. – PUNCT
Однажды – ADV|Degree=Pos
вечером – NOUN|Animacy=Inan|Case=Ins|Gender=Masc|Number=Sing
к – ADP
ней – PRON|Case=Dat|Gender=Fem|Number=Sing|Person=3
подошла – VERB|Aspect=Perf|Gender=Fem|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|V
oice=Act
девочка – NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing
с – ADP
велосипедом – NOUN|Animacy=Inan|Case=Ins|Gender=Masc|Number=Sing
и – CCONJ
спросила – VERB|Aspect=Perf|Gender=Fem|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|
Voice=Act
, – PUNCT
не – PART|Polarity=Neg
видела – VERB|Aspect=Imp|Gender=Fem|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voi
ce=Act
ли – PART
она – PRON|Case=Nom|Gender=Fem|Number=Sing|Person=3
ее – PRON|Case=Acc|Gender=Fem|Number=Sing|Person=3
пропавшего – VERB|Aspect=Perf|Case=Gen|Gender=Masc|Number=Sing|Tense=Past|VerbForm=P
art|Voice=Act
котенка – NOUN|Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing
. – PUNCT
```

Out[43]:
```
[None, None, None, None]
```

Лемматизация

Ввод [45]:
```
# Spacy
for token in spacy_text1:
      print(token, token.lemma, token.lemma_)
```

```
В 15939375860797385675 в
парке 692657576553750008 парк
на 16191904166009283104 на
краю 980890529103078125 край
города 6063391427805833384 город
растут 10901876232466620837 расти
старые 4368933178171963056 старый
дубы 11860818977535489124 дубы
и 15015917632809974589 и
ясени 5241734277658022696 ясени
. 12646065887601541794 .
Каждый 86315492416239735 00 каждый
вечер 87199569610644303 79 вечер
туда 12581053013947701146 туда
приходит 54101402717735293 33 приходить
пожилая 144741215843147915 61 пожилой
женщина 147118303648649770 40 женщина
с 5863529159893111856 с
белой 8754127145112654094 белый
собакой 171908850730723903 35 собака
. 12646065887601541794 .
Она 16361232442883285 34 она
любит 3482015854755669259 любить
сидеть 8581546620515527494 сидеть
на 16191904166009283104 на
скамейке 107580947897595436 90 скамейка
и 15015917632809974589 и
читать 165599639861631611 2 читать
книги 6587368699761570485 книга
, 2593208677638477497 ,
пока 33440207908666146 58 пока
собака 171908850730723903 35 собака
бегает 1325073068706028034 бегать
по 12047934663327436226 по
лужайке 6668165599393936598 лужайка
. 12646065887601541794 .
Однажды 5035183580430896573 однажды
вечером 87199569610644303 79 вечер
к 2390146911029080849 к
ней 6370563992700638067 ней
подошла 163450888767909659400 подойти
девочка 15450349728185719425 девочка
с 5863529159893111856 с
велосипедом 457832749362563597 велосипед
и 15015917632809974589 и
спросила 11529464371375940213 спросить
, 2593208677638477497 ,
не 5319710824202933802 не
видела 113855729893872883 87 видеть
ли 16253106445386410 77 ли
она 16361232442883285 34 она
ее 16361232442883285 34 она
пропавшего 1107468537556225995 пропасть
котенка 8384255308170038330 котёнок
. 12646065887601541794 .
```

Ввод [48]:
```
# Natasha
! pip install natasha
```

```
Collecting natasha
  Obtaining dependency information for natasha from https://files.pythonhosted.org/packages/32/9c/bb9d33c13564bcc939bb727087ef51b16ed3b49cc3b8fdec07c87b02f1de/natasha-1.6.0-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/32/9c/bb9d33c13564bcc939bb727087ef51b16ed3b49cc3b8fdec07c87b02f1de/natasha-1.6.0-py3-none-any.whl.metadata)
  Downloading natasha-1.6.0-py3-none-any.whl.metadata (23 kB)
Collecting pymorphy2 (from natasha)
  Obtaining dependency information for pymorphy2 from https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9886b64a54b476e1a332c67eee9f88e7f1ae8c9/pymorphy2-0.9.1-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9886b64a54b476e1a332c67eee9f88e7f1ae8c9/pymorphy2-0.9.1-py3-none-any.whl.metadata)
  Downloading pymorphy2-0.9.1-py3-none-any.whl.metadata (3.6 kB)
Requirement already satisfied: razdel>=0.5.0 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from natasha) (0.5.0)
Requirement already satisfied: navec>=0.9.0 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from natasha) (0.10.0)
Requirement already satisfied: slovnet>=0.6.0 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from natasha) (0.6.0)
Collecting yargy>=0.16.0 (from natasha)
  Obtaining dependency information for yargy>=0.16.0 from https://files.pythonhosted.org/packages/b7/55/d065a9812c619889fbe01a1863743ee45f7c60c462fc95b19576972ee9e4/yargy-0.16.0-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/b7/55/d065a9812c619889fbe01a1863743ee45f7c60c462fc95b19576972ee9e4/yargy-0.16.0-py3-none-any.whl.metadata)
  Downloading yargy-0.16.0-py3-none-any.whl.metadata (3.5 kB)
Collecting ipymarkup>=0.8.0 (from natasha)
  Obtaining dependency information for ipymarkup>=0.8.0 from https://files.pythonhosted.org/packages/bf/9b/bf54c98d50735a4a7c84c71e92c5361730c878ebfe903d2c2d196ef66055/ipymarkup-0.9.0-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/bf/9b/bf54c98d50735a4a7c84c71e92c5361730c878ebfe903d2c2d196ef66055/ipymarkup-0.9.0-py3-none-any.whl.metadata)
  Downloading ipymarkup-0.9.0-py3-none-any.whl.metadata (5.6 kB)
Requirement already satisfied: intervaltree>=3 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from ipymarkup>=0.8.0->natasha) (3.1.0)
Requirement already satisfied: numpy in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from navec>=0.9.0->natasha) (1.26.4)
Requirement already satisfied: dawg-python>=0.7.1 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from pymorphy2->natasha) (0.7.2)
Collecting pymorphy2-dicts-ru<3.0,>=2.4 (from pymorphy2->natasha)
  Obtaining dependency information for pymorphy2-dicts-ru<3.0,>=2.4 from https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9e96badf174068a2dd20264b9a378f2be1cdd9e/pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9e96badf174068a2dd20264b9a378f2be1cdd9e/pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl.metadata)
  Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl.metadata (2.1 kB)
Collecting docopt>=0.6 (from pymorphy2->natasha)
  Downloading docopt-0.6.2.tar.gz (25 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: sortedcontainers<3.0,>=2.0 in /Users/evseykirichkov/anaconda3/lib/python3.11/site-packages (from intervaltree>=3->ipymarkup>=0.8.0->natasha) (2.4.0)
Downloading natasha-1.6.0-py3-none-any.whl (34.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 34.4/34.4 MB 5.2 MB/s eta 0:00:0000:0100:01m
Downloading ipymarkup-0.9.0-py3-none-any.whl (14 kB)
Downloading yargy-0.16.0-py3-none-any.whl (33 kB)
Downloading pymorphy2-0.9.1-py3-none-any.whl (55 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.5/55.5 kB 4.8 MB/s eta 0:00:00
Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl (8.2 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.2/8.2 MB 5.8 MB/s eta 0:00:0000:0100:01
Building wheels for collected packages: docopt
  Building wheel for docopt (setup.py) ... done
  Created wheel for docopt: filename=docopt-0.6.2-py2.py3-none-any.whl size=13705 sha256=297e670ffb0f38dac888648f2703975938765d8856eb177e7340f917002d4fa0
  Stored in directory: /Users/evseykirichkov/Library/Caches/pip/wheels/1a/b0/8c/4b75
```

```
  c4116c31f83c8f9f047231251e13cc74481cca4a78a9ce
  Successfully built docopt
  Installing collected packages: pymorphy2-dicts-ru, docopt, pymorphy2, yargy, ipymark
  up, natasha
  Successfully installed docopt-0.6.2 ipymarkup-0.9.0 natasha-1.6.0 pymorphy2-0.9.1 py
  morphy2-dicts-ru-2.4.417127.4579844 yargy-0.16.0
```

Ввод [49]:
```
from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, MorphVocab
```

Ввод [50]:
```
def n_lemmatize(text):
    emb = NewsEmbedding()
    morph_tagger = NewsMorphTagger(emb)
    segmenter = Segmenter()
    morph_vocab = MorphVocab()
    doc = Doc(text)
    doc.segment(segmenter)
    doc.tag_morph(morph_tagger)
    for token in doc.tokens:
        token.lemmatize(morph_vocab)
    return doc
```

Ввод [51]:
```
n_doc1 = n_lemmatize(text1)
{_.text: _.lemma for _ in n_doc1.tokens}
```

Out[51]:
```
{'В': 'в',
 'парке': 'парк',
 'на': 'на',
 'краю': 'край',
 'города': 'город',
 'растут': 'расти',
 'старые': 'старый',
 'дубы': 'дуб',
 'и': 'и',
 'ясени': 'ясень',
 '.': '.',
 'Каждый': 'каждый',
 'вечер': 'вечер',
 'туда': 'туда',
 'приходит': 'приходить',
 'пожилая': 'пожилой',
 'женщина': 'женщина',
 'с': 'с',
 'белой': 'белый',
 'собакой': 'собака',
 'Она': 'она',
 'любит': 'любить',
 'сидеть': 'сидеть',
 'скамейке': 'скамейка',
 'читать': 'читать',
 'книги': 'книга',
 ',': ',',
 'пока': 'пока',
 'собака': 'собака',
 'бегает': 'бегать',
 'по': 'по',
 'лужайке': 'лужайка',
 'Однажды': 'однажды',
 'вечером': 'вечер',
 'к': 'к',
 'ней': 'она',
 'подошла': 'подойти',
 'девочка': 'девочка',
 'велосипедом': 'велосипед',
 'спросила': 'спросить',
 'не': 'не',
 'видела': 'видеть',
 'ли': 'ли',
 'она': 'она',
 'ее': 'она',
 'пропавшего': 'пропасть',
 'котенка': 'котенок'}
```

Выделение (распознавание) именованных сущностей

Ввод [58]:
```
# Spacy
for ent in spacy_text2.ents:
    print(ent.text, ent.label_)
```

```
Мария PER
```

Ввод [61]:
```
from spacy import displacy
displacy.render(spacy_text2, style='ent', jupyter=True)
```

В офисе всегда царит суета по утрам. Сотрудники спешат к своим рабочим местам, на столах дымятся чашки с горячим кофе, и слышатся звуки печатающих клавиатур. Мария **PER** , новая сотрудница, нервно листала страницы своего блокнота, готовясь к первой встрече с клиентами. Ее коллеги подбадривали ее, уверяя, что все пройдет отлично.

Ввод [63]:
```
print(spacy.explain("PER"))
```

```
Named person or family.
```

Ввод [64]:
```
# Natasha
from slovnet import NER
from ipymarkup import show_span_ascii_markup as show_markup
```

Ввод [69]:
```
ner = NER.load('data/slovnet_ner_news_v1.tar')
ner_res = ner.navec(navec)
markup_ner3 = ner(text2)
markup_ner3
```

Out[69]:
```
SpanMarkup(
    text='В офисе всегда царит суета по утрам. Сотрудники спешат к своим рабочим мес
там, на столах дымятся чашки с горячим кофе, и слышатся звуки печатающих клавиатур.
Мария, новая сотрудница, нервно листала страницы своего блокнота, готовясь к первой
встрече с клиентами. Ее коллеги подбадривали ее, уверяя, что все пройдет отлично.',
    spans=[Span(
         start=158,
         stop=163,
         type='PER'
     )]
)
```

Ввод [70]:
```
show_markup(markup_ner3.text, markup_ner3.spans)
```

```
В офисе всегда царит суета по утрам. Сотрудники спешат к своим рабочим
 местам, на столах дымятся чашки с горячим кофе, и слышатся звуки
печатающих клавиатур. Мария, новая сотрудница, нервно листала страницы
                      PER——
 своего блокнота, готовясь к первой встрече с клиентами. Ее коллеги
подбадривали ее, уверяя, что все пройдет отлично.
```

Разбор предложения

Ввод [71]:
```
# Spacy
from spacy import displacy
```

Ввод [72]:
```
displacy.render(spacy_text1, style='dep', jupyter=True)
```

В ADP парке NOUN на ADP краю NOUN города NOUN растут VERB старые ADJ дубы NOUN и CCONJ ясени. NOUN Каждый DET вечер NOUN туда ADV приходит VERB пожилая ADJ женщина NOUN с ADP белой ADJ собакой. NOUN Она PRON любит VERB сидеть VERB на ADP скамейке NOUN и CCONJ читать VERB книги, NOUN пока SCONJ собака NOUN бегает VERB по ADP лужайке. NOUN Однажды ADV вечером NOUN к ADP ней PRON подошла VERB девочка NOUN с ADP велосипедом NOUN и CCONJ спросила, VERB не PART видела VERB ли PART она PRON ее DET пропавшего VERB котенка. NOUN case obl case nmod nmod amod nsubj cc conj det obl advmod amod nsubj case amod nmod nsubj xcomp case obl cc conj obj mark nsubj advcl case obl advmod obl case obl nsubj case nmod cc conj advmod ccomp advmod nsubj det amod obj

Ввод [74]:
```
print(spacy.explain("ADP"))
```

```
adposition
```

Ввод [76]:
```
# Natasha
from natasha import NewsSyntaxParser
```

Ввод [79]:
```
emb = NewsEmbedding()
syntax_parser = NewsSyntaxParser(emb)
n_doc1.parse_syntax(syntax_parser)
n_doc1.sents[0].syntax.print()
```

```
        ┌─► В      case
   ┌─►┌─┴─ парке  obl
   │  │  ┌─► на     case
   │  └─►└─ краю   nmod
   │     └─► города nmod
┌──┴──┬──── растут
│     │  ┌─► старые amod
│     └─►└─ дубы   nsubj
│        │  ┌─► и      cc
│        └─►└─ ясени  conj
└─────────► .      punct
```

Ввод [80]:
```
n_doc1.parse_syntax(syntax_parser)
n_doc1.sents[1].syntax.print()
```

```
           ┌─► Каждый   det
        ┌─►┌─┴─ вечер    obl
        │  │  ┌─► туда     advmod
        │  └─►└─ приходит
┌───────┴──┬──── пожилая  amod
│          │  ┌─► женщина  nsubj
│          │  │  ┌─► с        case
│          │  │  ├─► белой    amod
│          └─►└─►└─ собакой  nmod
└────────────────► .        punct
```

Итог

Продемонстрировано использование NLTK, Spacy и Natasha.

Ввод [ ]:
```
```