Лабораторная работа №5

Предобработка текста.

Для произвольного предложения или текста решите следующие задачи:

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

In [1]:
```
text = """
Токенизация (иногда — сегментация) по предложениям — это процесс разделения письменно
Идея выглядит довольно простой. В английском и некоторых других языках мы можем вы
находим определенный знак пунктуации — точку.
"""
```

Токенизация

In [3]:
```
!pip install -U nltk
!pip install -U spacy
!python -m spacy download ru_core_news_sm
```

```
Collecting nltk
  Downloading nltk-3.6.2-py3-none-any.whl (1.5 MB)
     |████████████████████████████████| 1.5 MB 1.3 MB/s eta 0:00:01
Collecting tqdm
  Downloading tqdm-4.60.0-py2.py3-none-any.whl (75 kB)
     |████████████████████████████████| 75 kB 3.3 MB/s eta 0:00:011
Requirement already satisfied: joblib in /Users/ruapanc/virtualenvs/pis/
lib/python3.7/site-packages (from nltk) (1.0.1)
Collecting click
  Downloading click-8.0.1-py3-none-any.whl (97 kB)
     |████████████████████████████████| 97 kB 2.3 MB/s eta 0:00:011
Collecting regex
  Downloading regex-2021.4.4-cp37-cp37m-macosx_10_9_x86_64.whl (285 kB)
     |████████████████████████████████| 285 kB 2.2 MB/s eta 0:00:01
Requirement already satisfied: importlib-metadata in /Users/ruapanc/virt
ualenvs/pis/lib/python3.7/site-packages (from click->nltk) (3.4.0)
Requirement already satisfied: typing-extensions>=3.6.4 in /Users/ruapan
c/virtualenvs/pis/lib/python3.7/site-packages (from importlib-metadata->
click->nltk) (3.7.4.3)
Requirement already satisfied: zipp>=0.5 in /Users/ruapanc/virtualenvs/p
is/lib/python3.7/site-packages (from importlib-metadata->click->nltk)
(3.4.0)
Installing collected packages: tqdm, regex, click, nltk
Successfully installed click-8.0.1 nltk-3.6.2 regex-2021.4.4 tqdm-4.60.0
Collecting spacy
  Downloading spacy-3.0.6-cp37-cp37m-macosx_10_9_x86_64.whl (12.4 MB)
     |████████████████████████████████| 12.4 MB 4.4 MB/s eta 0:00:01
Requirement already satisfied: setuptools in /Users/ruapanc/virtualenvs/
pis/lib/python3.7/site-packages (from spacy) (52.0.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/ruapanc/vir
```

```
tualenvs/pis/lib/python3.7/site-packages (from spacy) (4.60.0)
Collecting catalogue<2.1.0,>=2.0.3
  Downloading catalogue-2.0.4-py3-none-any.whl (16 kB)
Collecting srsly<3.0.0,>=2.4.1
  Downloading srsly-2.4.1-cp37-cp37m-macosx_10_9_x86_64.whl (449 kB)
     |████████████████████████████████| 449 kB 11.3 MB/s eta 0:00:01
Collecting requests<3.0.0,>=2.13.0
  Downloading requests-2.25.1-py2.py3-none-any.whl (61 kB)
     |████████████████████████████████| 61 kB 15.7 MB/s eta 0:00:01
Requirement already satisfied: jinja2 in /Users/ruapanc/virtualenvs/pis/
lib/python3.7/site-packages (from spacy) (2.11.3)
Requirement already satisfied: packaging>=20.0 in /Users/ruapanc/virtual
envs/pis/lib/python3.7/site-packages (from spacy) (20.9)
Collecting typer<0.4.0,>=0.3.0
  Downloading typer-0.3.2-py3-none-any.whl (21 kB)
Collecting thinc<8.1.0,>=8.0.3
  Downloading thinc-8.0.3-cp37-cp37m-macosx_10_9_x86_64.whl (1.1 MB)
     |████████████████████████████████| 1.1 MB 49.5 MB/s eta 0:00:01
Requirement already satisfied: typing-extensions<4.0.0.0,>=3.7.4 in /Use
rs/ruapanc/virtualenvs/pis/lib/python3.7/site-packages (from spacy) (3.
7.4.3)
Collecting cymem<2.1.0,>=2.0.2
  Downloading cymem-2.0.5-cp37-cp37m-macosx_10_9_x86_64.whl (31 kB)
Requirement already satisfied: numpy>=1.15.0 in /Users/ruapanc/virtualen
vs/pis/lib/python3.7/site-packages (from spacy) (1.18.5)
Collecting pydantic<1.8.0,>=1.7.1
  Downloading pydantic-1.7.4-cp37-cp37m-macosx_10_9_x86_64.whl (2.3 MB)
     |████████████████████████████████| 2.3 MB 39.1 MB/s eta 0:00:01
Collecting murmurhash<1.1.0,>=0.28.0
  Downloading murmurhash-1.0.5-cp37-cp37m-macosx_10_9_x86_64.whl (18 kB)
Collecting blis<0.8.0,>=0.4.0
  Downloading blis-0.7.4-cp37-cp37m-macosx_10_9_x86_64.whl (5.8 MB)
     |████████████████████████████████| 5.8 MB 4.1 MB/s eta 0:00:01
Collecting preshed<3.1.0,>=3.0.2
  Downloading preshed-3.0.5-cp37-cp37m-macosx_10_9_x86_64.whl (104 kB)
     |████████████████████████████████| 104 kB 42.7 MB/s eta 0:00:01
Collecting wasabi<1.1.0,>=0.8.1
  Downloading wasabi-0.8.2-py3-none-any.whl (23 kB)
Collecting pathy>=0.3.5
  Downloading pathy-0.5.2-py3-none-any.whl (42 kB)
     |████████████████████████████████| 42 kB 4.2 MB/s  eta 0:00:01
Collecting spacy-legacy<3.1.0,>=3.0.4
  Downloading spacy_legacy-3.0.5-py2.py3-none-any.whl (12 kB)
Requirement already satisfied: zipp>=0.5 in /Users/ruapanc/virtualenvs/p
is/lib/python3.7/site-packages (from catalogue<2.1.0,>=2.0.3->spacy) (3.
4.0)
Requirement already satisfied: pyparsing>=2.0.2 in /Users/ruapanc/virtua
lenvs/pis/lib/python3.7/site-packages (from packaging>=20.0->spacy) (2.
4.7)
Collecting smart-open<4.0.0,>=2.2.0
  Downloading smart_open-3.0.0.tar.gz (113 kB)
     |████████████████████████████████| 113 kB 28.9 MB/s eta 0:00:01
Collecting idna<3,>=2.5
  Using cached idna-2.10-py2.py3-none-any.whl (58 kB)
Collecting certifi>=2017.4.17
  Downloading certifi-2020.12.5-py2.py3-none-any.whl (147 kB)
     |████████████████████████████████| 147 kB 8.7 MB/s eta 0:00:01
Collecting chardet<5,>=3.0.2
  Downloading chardet-4.0.0-py2.py3-none-any.whl (178 kB)
     |████████████████████████████████| 178 kB 39.1 MB/s eta 0:00:01
Collecting urllib3<1.27,>=1.21.1
  Downloading urllib3-1.26.4-py2.py3-none-any.whl (153 kB)
     |████████████████████████████████| 153 kB 32.5 MB/s eta 0:00:01
Collecting click<7.2.0,>=7.1.1
  Using cached click-7.1.2-py2.py3-none-any.whl (82 kB)
Requirement already satisfied: MarkupSafe>=0.23 in /Users/ruapanc/virtua
lenvs/pis/lib/python3.7/site-packages (from jinja2->spacy) (1.1.1)
Building wheels for collected packages: smart-open
  Building wheel for smart-open (setup.py) ... done
```

```
  Created wheel for smart-open: filename=smart_open-3.0.0-py3-none-any.w
hl size=107097 sha256=a578d86f02703b313fdb45f39fd08dcd82221f9086027e8869
2c2311eb5da3cc
  Stored in directory: /Users/ruapanc/Library/Caches/pip/wheels/83/a6/1
2/bf3c1a667bde4251be5b7a3368b2d604c9af2105b5c1cb1870
Successfully built smart-open
Installing collected packages: urllib3, idna, chardet, certifi, request
s, murmurhash, cymem, click, catalogue, wasabi, typer, srsly, smart-ope
n, pydantic, preshed, blis, thinc, spacy-legacy, pathy, spacy
  Attempting uninstall: click
    Found existing installation: click 8.0.1
    Uninstalling click-8.0.1:
      Successfully uninstalled click-8.0.1
Successfully installed blis-0.7.4 catalogue-2.0.4 certifi-2020.12.5 char
det-4.0.0 click-7.1.2 cymem-2.0.5 idna-2.10 murmurhash-1.0.5 pathy-0.5.2
preshed-3.0.5 pydantic-1.7.4 requests-2.25.1 smart-open-3.0.0 spacy-3.0.
6 spacy-legacy-3.0.5 srsly-2.4.1 thinc-8.0.3 typer-0.3.2 urllib3-1.26.4
wasabi-0.8.2
Collecting ru-core-news-sm==3.0.0
  Downloading https://github.com/explosion/spacy-models/releases/downloa
d/ru_core_news_sm-3.0.0/ru_core_news_sm-3.0.0-py3-none-any.whl (17.9 MB)
     |████████████████████████████████| 17.9 MB 9.7 MB/s eta 0:00:01
Requirement already satisfied: spacy<3.1.0,>=3.0.0 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from ru-core-news-sm==3.0.0)
(3.0.6)
Collecting pymorphy2>=0.9
  Downloading pymorphy2-0.9.1-py3-none-any.whl (55 kB)
     |████████████████████████████████| 55 kB 1.6 MB/s eta 0:00:01
Collecting dawg-python>=0.7.1
  Downloading DAWG_Python-0.7.2-py2.py3-none-any.whl (11 kB)
Collecting pymorphy2-dicts-ru<3.0,>=2.4
  Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl
(8.2 MB)
     |████████████████████████████████| 8.2 MB 3.8 MB/s eta 0:00:01
Collecting docopt>=0.6
  Downloading docopt-0.6.2.tar.gz (25 kB)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-c
ore-news-sm==3.0.0) (2.0.5)
Requirement already satisfied: pydantic<1.8.0,>=1.7.1 in /Users/ruapanc/
virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->r
u-core-news-sm==3.0.0) (1.7.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/ruapa
nc/virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0
->ru-core-news-sm==3.0.0) (1.0.5)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in /Users/ruapanc/virt
ualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-co
re-news-sm==3.0.0) (0.7.4)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-c
ore-news-sm==3.0.0) (2.4.1)
Requirement already satisfied: thinc<8.1.0,>=8.0.3 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-c
ore-news-sm==3.0.0) (8.0.3)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in /Users/ruapanc/vi
rtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-
core-news-sm==3.0.0) (0.8.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.4 in /Users/ruap
anc/virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.
0->ru-core-news-sm==3.0.0) (3.0.5)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/ruapanc/v
irtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru
-core-news-sm==3.0.0) (3.0.5)
Requirement already satisfied: jinja2 in /Users/ruapanc/virtualenvs/pis/
lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-core-news-sm==
3.0.0) (2.11.3)
Requirement already satisfied: numpy>=1.15.0 in /Users/ruapanc/virtualen
vs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-core-ne
ws-sm==3.0.0) (1.18.5)
```

```
Requirement already satisfied: catalogue<2.1.0,>=2.0.3 in /Users/ruapan
c/virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0-
>ru-core-news-sm==3.0.0) (2.0.4)
Requirement already satisfied: packaging>=20.0 in /Users/ruapanc/virtual
envs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-core-
news-sm==3.0.0) (20.9)
Requirement already satisfied: pathy>=0.3.5 in /Users/ruapanc/virtualenv
s/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-core-new
s-sm==3.0.0) (0.5.2)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-c
ore-news-sm==3.0.0) (4.60.0)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /Users/ruapan
c/virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0-
>ru-core-news-sm==3.0.0) (2.25.1)
Requirement already satisfied: typer<0.4.0,>=0.3.0 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-c
ore-news-sm==3.0.0) (0.3.2)
Requirement already satisfied: setuptools in /Users/ruapanc/virtualenvs/
pis/lib/python3.7/site-packages (from spacy<3.1.0,>=3.0.0->ru-core-news-
sm==3.0.0) (52.0.0)
Requirement already satisfied: typing-extensions<4.0.0.0,>=3.7.4 in /Use
rs/ruapanc/virtualenvs/pis/lib/python3.7/site-packages (from spacy<3.1.
0,>=3.0.0->ru-core-news-sm==3.0.0) (3.7.4.3)
Requirement already satisfied: zipp>=0.5 in /Users/ruapanc/virtualenvs/p
is/lib/python3.7/site-packages (from catalogue<2.1.0,>=2.0.3->spacy<3.1.
0,>=3.0.0->ru-core-news-sm==3.0.0) (3.4.0)
Requirement already satisfied: pyparsing>=2.0.2 in /Users/ruapanc/virtua
lenvs/pis/lib/python3.7/site-packages (from packaging>=20.0->spacy<3.1.
0,>=3.0.0->ru-core-news-sm==3.0.0) (2.4.7)
Requirement already satisfied: smart-open<4.0.0,>=2.2.0 in /Users/ruapan
c/virtualenvs/pis/lib/python3.7/site-packages (from pathy>=0.3.5->spacy<
3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (3.0.0)
Requirement already satisfied: certifi>=2017.4.17 in /Users/ruapanc/virt
ualenvs/pis/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0->s
pacy<3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (2020.12.5)
Requirement already satisfied: chardet<5,>=3.0.2 in /Users/ruapanc/virtu
alenvs/pis/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0->sp
acy<3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (4.0.0)
Requirement already satisfied: idna<3,>=2.5 in /Users/ruapanc/virtualenv
s/pis/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0->spacy<
3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/ruapanc/v
irtualenvs/pis/lib/python3.7/site-packages (from requests<3.0.0,>=2.13.0
->spacy<3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (1.26.4)
Requirement already satisfied: click<7.2.0,>=7.1.1 in /Users/ruapanc/vir
tualenvs/pis/lib/python3.7/site-packages (from typer<0.4.0,>=0.3.0->spac
y<3.1.0,>=3.0.0->ru-core-news-sm==3.0.0) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in /Users/ruapanc/virtua
lenvs/pis/lib/python3.7/site-packages (from jinja2->spacy<3.1.0,>=3.0.0-
>ru-core-news-sm==3.0.0) (1.1.1)
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 s
ize=13705 sha256=fbb26eaca161cbf1af7c73a419e9834ee7acaa2982154219d807041
2f5a7c5e9
  Stored in directory: /Users/ruapanc/Library/Caches/pip/wheels/72/b0/3
f/1d95f96ff986c7dfffe46ce2be4062f38ebd04b506c77c81b9
Successfully built docopt
Installing collected packages: pymorphy2-dicts-ru, docopt, dawg-python,
pymorphy2, ru-core-news-sm
Successfully installed dawg-python-0.7.2 docopt-0.6.2 pymorphy2-0.9.1 py
morphy2-dicts-ru-2.4.417127.4579844 ru-core-news-sm-3.0.0
✔ Download and installation successful
You can now load the package via spacy.load('ru_core_news_sm')
```

In [4]:

```
import nltk
from nltk import tokenize
```

```
nltk_tk_1 = nltk.WordPunctTokenizer()
nltk_tk_1.tokenize(text)
```

Out[4]:
```
['Токенизация',
 '(',
 'иногда',
 '—',
 'сегментация',
 ')',
 'по',
 'предложениям',
 '—',
 'это',
 'процесс',
 'разделения',
 'письменного',
 'языка',
 'на',
 'предложения',
 '-',
 'компоненты',
 '.',
 'Идея',
 'выглядит',
 'довольно',
 'простой',
 '.',
 'В',
 'английском',
 'и',
 'некоторых',
 'других',
 'языках',
 'мы',
 'можем',
 'вычленять',
 'предложение',
 'каждый',
 'раз',
 ',',
 'когда',
 'находим',
 'определенный',
 'знак',
 'пунктуации',
 '—',
 'точку',
 '.']
```

In [8]:
```
!pip install -U razdel
```

```
Collecting razdel
  Downloading razdel-0.5.0-py3-none-any.whl (21 kB)
Installing collected packages: razdel
Successfully installed razdel-0.5.0
```

In [9]:
```
from razdel import tokenize, sentenize
n_tok_text1 = list(tokenize(text))
n_tok_text1
```

Out[9]:
```
[Substring(1, 12, 'Токенизация'),
 Substring(13, 14, '('),
 Substring(14, 20, 'иногда'),
 Substring(21, 22, '—'),
 Substring(23, 34, 'сегментация'),
 Substring(34, 35, ')'),
```

```
 Substring(36, 38, 'по'),
 Substring(39, 51, 'предложениям'),
 Substring(52, 53, '—'),
 Substring(54, 57, 'это'),
 Substring(58, 65, 'процесс'),
 Substring(66, 76, 'разделения'),
 Substring(77, 88, 'письменного'),
 Substring(89, 94, 'языка'),
 Substring(95, 97, 'на'),
 Substring(98, 120, 'предложения-компоненты'),
 Substring(120, 121, '.'),
 Substring(123, 127, 'Идея'),
 Substring(128, 136, 'выглядит'),
 Substring(137, 145, 'довольно'),
 Substring(146, 153, 'простой'),
 Substring(153, 154, '.'),
 Substring(155, 156, 'В'),
 Substring(157, 167, 'английском'),
 Substring(168, 169, 'и'),
 Substring(170, 179, 'некоторых'),
 Substring(180, 186, 'других'),
 Substring(187, 193, 'языках'),
 Substring(194, 196, 'мы'),
 Substring(197, 202, 'можем'),
 Substring(203, 212, 'вычленять'),
 Substring(213, 224, 'предложение'),
 Substring(225, 231, 'каждый'),
 Substring(232, 235, 'раз'),
 Substring(235, 236, ','),
 Substring(237, 242, 'когда'),
 Substring(244, 251, 'находим'),
 Substring(252, 264, 'определенный'),
 Substring(265, 269, 'знак'),
 Substring(270, 280, 'пунктуации'),
 Substring(281, 282, '—'),
 Substring(283, 288, 'точку'),
 Substring(288, 289, '.')]
```

In [10]:

```
list(sentenize(text))
```

Out[10]:

```
[Substring(1,
           121,
           'Токенизация (иногда — сегментация) по предложениям — это процесс раздел
ения письменного языка на предложения-компоненты.'),
 Substring(123, 154, 'Идея выглядит довольно простой.'),
 Substring(155,
           289,
           'В английском и некоторых других языках мы можем вычленять предложение
каждый раз, когда \nнаходим определенный знак пунктуации — точку.')]
```

In [6]:

```
from spacy.lang.ru import Russian
import spacy
nlp = spacy.load('ru_core_news_sm')
spacy_text1 = nlp(text)
spacy_text1
```

Out[6]:

Токенизация (иногда — сегментация) по предложениям — это процесс разделения письменного языка на предложения-компоненты.
Идея выглядит довольно простой. В английском и некоторых других языках мы можем вычленять предложение каждый раз, когда
находим определенный знак пунктуации — точку.

In [7]:

```
for t in spacy_text1:
    print(t)
```

Токенизация
(
иногда
—
сегментация
)
по
предложениям
—
это
процесс
разделения
письменного
языка
на
предложения
-
компоненты
.

Идея
выглядит
довольно
простой
.
В
английском
и
некоторых
других
языках
мы
можем
вычленять
предложение
каждый
раз
,
когда

находим
определенный
знак
пунктуации
—
точку
.

Частеречная разметка

In [11]:
```
for token in spacy_text1:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
 - SPACE - ROOT
Токенизация - PROPN - nsubj
( - PUNCT - punct
иногда - ADV - advmod
— - PUNCT - punct
сегментация - NOUN - appos
) - PUNCT - punct
по - ADP - case
предложениям - NOUN - nmod
— - PUNCT - punct
```

```
это - PART - expl
процесс - NOUN - ROOT
разделения - NOUN - nmod
письменного - ADJ - amod
языка - NOUN - nmod
на - ADP - case
предложения - NOUN - nmod
- - NOUN - nmod
компоненты - NOUN - nmod
. - PUNCT - punct

 - SPACE - ROOT
Идея - NOUN - nsubj
выглядит - VERB - ROOT
довольно - ADV - advmod
простой - ADJ - xcomp
. - PUNCT - punct
В - ADP - case
английском - ADJ - amod
и - CCONJ - cc
некоторых - DET - det
других - ADJ - amod
языках - NOUN - obl
мы - PRON - nsubj
можем - VERB - ROOT
вычленять - VERB - xcomp
предложение - NOUN - obj
каждый - DET - det
раз - NOUN - obl
, - PUNCT - punct
когда - ADV - mark

 - SPACE - discourse
находим - VERB - acl:relcl
определенный - ADJ - amod
знак - NOUN - obj
пунктуации - NOUN - nmod
— - PUNCT - punct
точку - NOUN - appos
. - PUNCT - punct

 - SPACE - punct
```

In [32]:

```
print(spacy.explain("ADJ"))
```

```
adjective
```

Лемматизация

In [18]:

```
for token in spacy_text1:
      print(token, token.lemma, token.lemma_)
```

```
 962983613142996970

Токенизация 7539190963720402728 токенизация
( 12638816674900267446 (
иногда 11473840011754297908 иногда
— 10118409446379451916 —
сегментация 16712449363547716331 сегментация
) 3842344029291005339 )
по 12047934663327436226 по
предложениям 3896560238297484731 предложение
— 10118409446379451916 —
это 1823958246850563701 это
процесс 14462777509019072512 процесс
разделения 10991016642072250773 разделение
```

```
письменного 2798609870607126203 письменный
языка 14510553211863083651 язык
на 16191904166009283104 на
предложения 3896560238297484731 предложение
– 9153284864653046197 –
компоненты 12090406748281175970 компонент
. 12646065887601541794 .

 962983613142996970

Идея 3551521432520947143 идея
выглядит 15801308832189027519 выглядеть
довольно 13336327071705512178 довольно
простой 17660191490670304307 простой
. 12646065887601541794 .
В 15939375860797385675 в
английском 15391208138942167357 английский
и 15015917632809974589 и
некоторых 2648419700197113123 некоторый
других 55685201222224931142 других
языках 14510553211863083651 язык
мы 82659241346168242 62 мы
можем 14329395112709808155 мочь
вычленять 3559133231479220877 вычленять
предложение 3896560238297484731 предложение
каждый 86315492416239735 00 каждый
раз 971133014553193710 раз
, 2593208677638477497 ,
когда 17611357251313260 45 когда

 962983613142996970

находим 13173750222482653110 находить
определенный 11016737793840018915 определённый
знак 14184539529555368913 знак
пунктуации 4758878851802152342 пунктуация
— 10118409446379451916 —
точку 11203916623887805 24 точка
. 12646065887601541794 .

 962983613142996970
```

In [21]:
```
from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, Morp
```

In [22]:
```
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
```

In [24]:
```
n_doc1 = n_lemmatize(text)
{_.text: _.lemma for _ in n_doc1.tokens}
```

Out[24]:
```
{'Токенизация': 'токенизация',
 '(': '(',
 'иногда': 'иногда',
 '—': '—',
```

```
'сегментация': 'сегментация',
')': ')',
'по': 'по',
'предложениям': 'предложение',
'это': 'это',
'процесс': 'процесс',
'разделения': 'разделение',
'письменного': 'письменный',
'языка': 'язык',
'на': 'на',
'предложения-компоненты': 'предложение-компонента',
'.': '.',
'Идея': 'идея',
'выглядит': 'выглядеть',
'довольно': 'довольно',
'простой': 'простой',
'В': 'в',
'английском': 'английский',
'и': 'и',
'некоторых': 'некоторый',
'других': 'другой',
'языках': 'язык',
'мы': 'мы',
'можем': 'мочь',
'вычленять': 'вычленять',
'предложение': 'предложение',
'каждый': 'каждый',
'раз': 'раз',
',': ',',
'когда': 'когда',
'находим': 'находить',
'определенный': 'определенный',
'знак': 'знак',
'пунктуации': 'пунктуация',
'точку': 'точка'}
```

In [ ]:

Выделение именованных сущностей

In [28]:
```
spacy_text3 = nlp('Станкевич Андрей Сергеевич — лауреат специальной премии корпор
for ent in spacy_text3.ents:
    print(ent.text, ent.label_)
```

```
Андрей Сергеевич PER
IBM ORG
```

In [39]:
```
displacy.serve(nlp('Станкевич Андрей Сергеевич — лауреат специальной премии корпо
```

Станкевич Андрей Сергеевич **PER** — лауреат специальной премии корпорации IBM **ORG** .

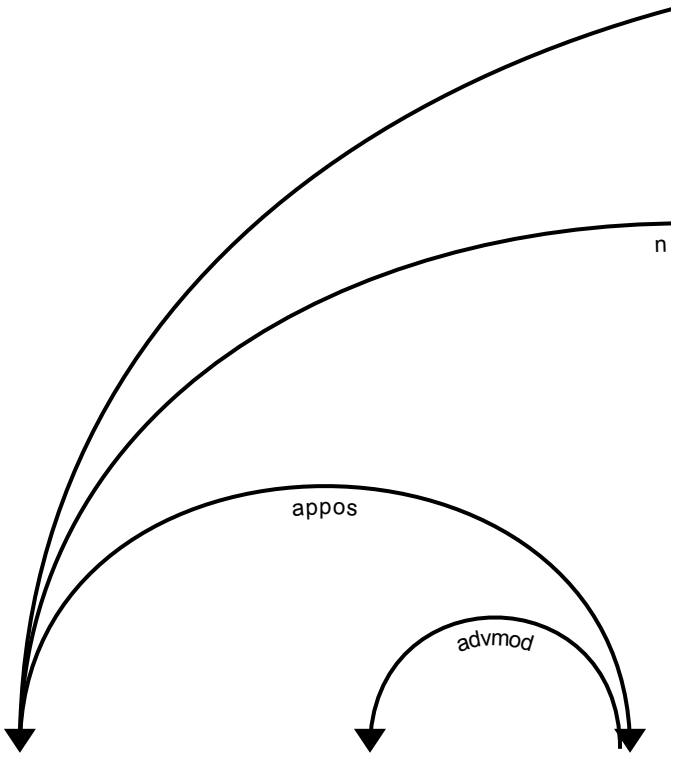
```
Using the 'ent' visualizer
Serving on http://0.0.0.0:5000 ...
```

```
127.0.0.1 - - [23/May/2021 18:42:04] "GET / HTTP/1.1" 200 1156
127.0.0.1 - - [23/May/2021 18:42:04] "GET /favicon.ico HTTP/1.1" 200 115
6
Shutting down server on port 5000.
```

Разбор предложения

In [29]:
```
from spacy import displacy
```

In [41]:
```
displacy.render(spacy_text1, style='dep', jupyter=True)
```



Токенизация (иногда – сегмен

SPACE PROPN ADV NC

In [ ]:
```

```