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 = """
Лемматизация – это метод морфологического анализа, который сводится к приведению сло
к ее первоначальной словарной форме (лемме). Метод лемматизации применяется в поиско
веб-документов при их индексировании.В результате лемматизации от словоформы отбрасы
основная или словарная форма слова.
"""
```

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 [2]:

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

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

Out[2]:
```
['Лемматизация',
 '–',
 'это',
 'метод',
 'морфологического',
 'анализа',
 ',',
 'который',
 'сводится',
 'к',
 'приведению',
 'словоформы',
 'к',
 'ее',
 'первоначальной',
 'словарной',
 'форме',
 '(',
 'лемме',
 ').',
 'Метод',
 'лемматизации',
 'применяется',
 'в',
 'поисковых',
 'алгоритмах',
 'в',
 'процессе',
 'схематизации',
 'веб',
 '–',
 'документов',
 'при',
 'их',
 'индексировании',
 '.',
 'В',
 'результате',
 'лемматизации',
 'от',
 'словоформы',
 'отбрасываются',
 'флективные',
 'окончания',
 'и',
 'возвращается',
 'основная',
 'или',
 'словарная',
 'форма',
 'слова',
 '.']
```

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 [3]:
```
from razdel import tokenize, sentenize
n_tok_text1 = list(tokenize(text))
n_tok_text1
```

Out[3]:
```
[Substring(1, 13, 'Лемматизация'),
 Substring(14, 15, '–'),
 Substring(16, 19, 'это'),
 Substring(20, 25, 'метод'),
 Substring(26, 42, 'морфологического'),
 Substring(43, 50, 'анализа'),
 Substring(50, 51, ','),
 Substring(52, 59, 'который'),
 Substring(60, 68, 'сводится'),
 Substring(69, 70, 'к'),
 Substring(71, 81, 'приведению'),
 Substring(82, 92, 'словоформы'),
 Substring(94, 95, 'к'),
 Substring(96, 98, 'ее'),
 Substring(99, 113, 'первоначальной'),
 Substring(114, 123, 'словарной'),
 Substring(124, 129, 'форме'),
 Substring(130, 131, '('),
 Substring(131, 136, 'лемме'),
 Substring(136, 137, ')'),
 Substring(137, 138, '.'),
 Substring(139, 144, 'Метод'),
 Substring(145, 157, 'лемматизации'),
 Substring(158, 169, 'применяется'),
 Substring(170, 171, 'в'),
 Substring(172, 181, 'поисковых'),
 Substring(182, 192, 'алгоритмах'),
 Substring(193, 194, 'в'),
 Substring(195, 203, 'процессе'),
 Substring(204, 216, 'схематизации'),
 Substring(218, 232, 'веб-документов'),
 Substring(233, 236, 'при'),
 Substring(237, 239, 'их'),
 Substring(240, 254, 'индексировании'),
 Substring(254, 255, '.'),
 Substring(255, 256, 'В'),
 Substring(257, 267, 'результате'),
 Substring(268, 280, 'лемматизации'),
 Substring(281, 283, 'от'),
 Substring(284, 294, 'словоформы'),
 Substring(295, 308, 'отбрасываются'),
 Substring(309, 319, 'флективные'),
 Substring(320, 329, 'окончания'),
 Substring(330, 331, 'и'),
 Substring(332, 344, 'возвращается'),
 Substring(346, 354, 'основная'),
 Substring(355, 358, 'или'),
 Substring(359, 368, 'словарная'),
 Substring(369, 374, 'форма'),
 Substring(375, 380, 'слова'),
 Substring(380, 381, '.')]
```

In [4]:
```
list(sentenize(text))
```

Out[4]:
```
[Substring(1,
           138,
           'Лемматизация – это метод морфологического анализа, который сводится к п
риведению словоформы \nк ее первоначальной словарной форме (лемме).'),
 Substring(139,
           381,
           'Метод лемматизации применяется в поисковых алгоритмах в процессе схемат
изации \nвеб-документов при их индексировании.В результате лемматизации от словоформы
отбрасываются флективные окончания и возвращается \nосновная или словарная форма слов
а.')]
```

In [5]:
```
from spacy.lang.ru import Russian
import spacy
```

```
nlp = spacy.load('ru_core_news_sm')
spacy_text1 = nlp(text)
spacy_text1
```

Out[5]: Лемматизация - это метод морфологического анализа, который сводится к приведению слов оформы
к ее первоначальной словарной форме (лемме). Метод лемматизации применяется в поиско вых алгоритмах в процессе схематизации
веб-документов при их индексировании.В результате лемматизации от словоформы отбрасыв аются флективные окончания и возвращается
основная или словарная форма слова.

In [6]:
```
for t in spacy_text1:
    print(t)
```

```
Лемматизация
-
это
метод
морфологического
анализа
,
который
сводится
к
приведению
словоформы


к
ее
первоначальной
словарной
форме
(
лемме
)
.
Метод
лемматизации
применяется
в
поисковых
алгоритмах
в
процессе
схематизации


веб
-
документов
при
их
индексировании
.
В
результате
лемматизации
от
словоформы
отбрасываются
флективные
окончания
и
возвращается
```

основная
или
словарная
форма
слова
.

Частеречная разметка

In [7]:
```
for token in spacy_text1:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
 - SPACE - ROOT
Лемматизация - NOUN - nsubj
- - PUNCT - nsubj
это - PART - nsubj
метод - NOUN - ROOT
морфологического - ADJ - amod
анализа - NOUN - nmod
, - PUNCT - punct
который - PRON - nsubj
сводится - VERB - acl:relcl
к - ADP - case
приведению - NOUN - obl
словоформы - NOUN - nmod

 - SPACE - appos
к - ADP - case
ее - DET - det
первоначальной - ADJ - amod
словарной - ADJ - amod
форме - NOUN - nmod
( - PUNCT - punct
лемме - NOUN - parataxis
) - PUNCT - punct
. - PUNCT - punct
Метод - NOUN - nsubj:pass
лемматизации - NOUN - nmod
применяется - VERB - ROOT
в - ADP - case
поисковых - ADJ - amod
алгоритмах - NOUN - obl
в - ADP - case
процессе - NOUN - obl
схематизации - NOUN - nmod

 - SPACE - appos
веб - NOUN - nmod
- - NOUN - nmod
документов - NOUN - nmod
при - ADP - case
их - PRON - nmod
индексировании - NOUN - nmod
. - PUNCT - punct
В - ADP - case
результате - NOUN - obl
лемматизации - NOUN - nmod
от - ADP - case
словоформы - NOUN - obl
отбрасываются - VERB - ROOT
флективные - ADJ - amod
окончания - NOUN - nsubj:pass
и - CCONJ - cc
возвращается - VERB - conj
```

```
 - SPACE - xcomp
основная - ADJ - amod
или - CCONJ - cc
словарная - ADJ - conj
форма - NOUN - nsubj
слова - NOUN - nmod
. - PUNCT - punct

 - SPACE - ROOT
```

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

```
conjunct
```

Лемматизация

In [10]:
```
for token in spacy_text1:
      print(token, token.lemma, token.lemma_)
```

```
 962983613142996970

Лемматизация 13762791536589156665 лемматизация
- 9153284864653046197 -
это 1823958246850563701 это
метод 14317775670369151318 метод
морфологического 10107527482202187288 морфологический
анализа 10328217384203977629 анализ
, 2593208677638477497 ,
который 1512420474390618213 который
сводится 4984490614007732285 сводиться
к 2390146911029080849 к
приведению 11644771811425043626 приведение
словоформы 5030668642616423873 словоформа

 962983613142996970

к 2390146911029080849 к
ее 1267933210776559268 её
первоначальной 9480244986737757179 первоначальный
словарной 11495571827166064762 словарный
форме 9210309821255511689 форма
( 12638816674900267446 (
лемме 16393141331216114713 лемма
) 3842344029291005339 )
. 12646065887601541794 .
Метод 14317775670369151318 метод
лемматизации 13762791536589156665 лемматизация
применяется 11828352846431681784 применяться
в 15939375860797385675 в
поисковых 2841176456001165844 поисковый
алгоритмах 12533358430471453281 алгоритм
в 15939375860797385675 в
процессе 14462777509019072512 процесс
схематизации 2753121182393499731 схематизация

 962983613142996970

веб 7677034528521278175 веб
- 9153284864653046197 -
документов 14641638662421594528 документов
при 4143642271851148588 при
их 5373766955765579565 их
индексировании 11556908596646085085 индексирование
. 12646065887601541794 .
В 15939375860797385675 в
```

```
результате 356081769442416127 результат
лемматизации 13762791536589156665 лемматизация
от 7547231311137123581 от
словоформы 5030668642616423873 словоформа
отбрасываются 5971040154481061656 отбрасываются
флективные 7746506564866685218 флективный
окончания 1263132673098777593 окончание
и 15015917632809974589 и
возвращается 5722393860007198 возвращаться

 962983613142996970

основная 10259815185512192724 основной
или 1530020831762146143 или
словарная 11495571827166064762 словарный
форма 9210309821255511689 форма
слова 1386213856741127517 слово
. 12646065887601541794 .

 962983613142996970
```

In [11]:
```
from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, Morp
```

In [12]:
```
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 [13]:
```
n_doc1 = n_lemmatize(text)
{_.text: _.lemma for _ in n_doc1.tokens}
```

Out[13]:
```
{'Лемматизация': 'лемматизация',
 '-': '-',
 'это': 'это',
 'метод': 'метод',
 'морфологического': 'морфологический',
 'анализа': 'анализ',
 ',': ',',
 'который': 'который',
 'сводится': 'сводиться',
 'к': 'к',
 'приведению': 'приведение',
 'словоформы': 'словоформа',
 'ее': 'ее',
 'первоначальной': 'первоначальный',
 'словарной': 'словарный',
 'форме': 'форма',
 '(': '(',
 'лемме': 'лемма',
 ')': ')',
 '.': '.',
 'Метод': 'метод',
 'лемматизации': 'лемматизация',
 'применяется': 'применяться',
 'в': 'в',
 'поисковых': 'поисковый',
```

```
'алгоритмах': 'алгоритм',
'процессе': 'процесс',
'схематизации': 'схематизация',
'веб-документов': 'веб-документ',
'при': 'при',
'их': 'их',
'индексировании': 'индексирование',
'В': 'в',
'результате': 'результат',
'от': 'от',
'отбрасываются': 'отбрасываются',
'флективные': 'флективный',
'окончания': 'окончание',
'и': 'и',
'возвращается': 'возвращаться',
'основная': 'основный',
'или': 'или',
'словарная': 'словарный',
'форма': 'форма',
'слова': 'слово'}
```

In [ ]:

Выделение именованных сущностей

In [14]:
```
per = 'Николай Эрнестович Бауман (1873—1905) — российский революционер, деятель
```

In [15]:
```
spacy_text3 = nlp(per)
for ent in spacy_text3.ents:
    print(ent.text, ent.label_)
```

```
Николай Эрнестович Бауман PER
РСДРП ORG
```

In [18]:
```
displacy.serve(nlp(per), style="ent")
```

```
/Users/ruapanc/virtualenvs/pis/lib/python3.7/site-packages/spacy/displac
y/__init__.py:97: UserWarning: [W011] It looks like you're calling displ
acy.serve from within a Jupyter notebook or a similar environment. This
likely means you're already running a local web server, so there's no ne
ed to make displaCy start another one. Instead, you should be able to re
place displacy.serve with displacy.render to show the visualization.
  warnings.warn(Warnings.W011)
```

Николай Эрнестович Бауман **PER** (1873—1905) — российский революционер, деятель большевистского крыла РСДРП **ORG** .

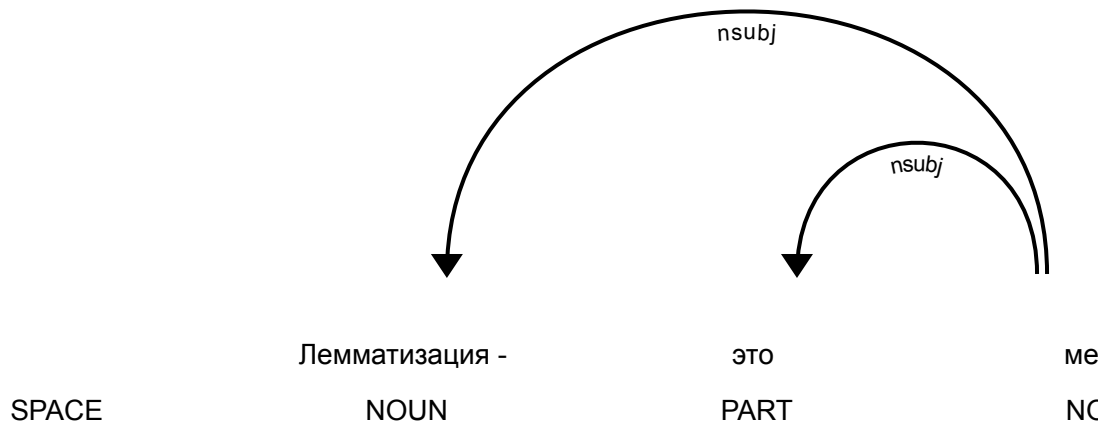
```
Using the 'ent' visualizer
Serving on http://0.0.0.0:5000 ...

Shutting down server on port 5000.
```

Разбор предложения

In [17]:
```
from spacy import displacy
```

In [19]:
```
displacy.render(spacy_text1, style='dep', jupyter=True)
```



In [ ]:
```

```