ЛРN°5 Предобработки текста

Чжан Чжибо ИУ5И-21М

Для выполнения работы испоьзована библиотека 'Natasha'

```
text='Сикорский родился 7 июня 1889 года. Он поступил в
```

Задача токенизации

```
!pip install razdel
from razdel import tokenize, sentenize
n_tok_text = list(tokenize(text))
n_tok_text
```

```
Collecting razdel
  Downloading https://files.pythonhosted.org/packages/15/2c/664223a3924aa6e70479f7d37220b3a65
Installing collected packages: razdel
Successfully installed razdel-0.5.0
[Substring(0, 9, 'Сикорский'),
 Substring(10, 17, 'родился'),
 Substring(18, 19, '7'),
 Substring(20, 24, 'июня'),
 Substring(25, 29, '1889'),
 Substring(30, 34, 'года'),
 Substring(34, 35, '.'),
 Substring(36, 38, 'Он'),
 Substring(39, 47, 'поступил'),
 Substring(48, 49, 'в'),
 Substring(50, 58, 'Киевский'),
 Substring(59, 74, 'политехнический'),
 Substring(75, 83, 'институт'),
 Substring(84, 85, 'в'),
 Substring(86, 90, '1907'),
 Substring(91, 95, 'году'),
 Substring(95, 96, '.'),
 Substring(97, 98, 'В'),
 Substring(99, 108, '1909-1912'),
 Substring(109, 114, 'года'),
 Substring(115, 122, 'студент'),
 Substring(123, 132, 'Сикорский'),
 Substring(133, 146, 'спроектировал'),
 Substring(147, 148, 'и'),
 Substring(149, 157, 'построил'),
 Substring(158, 161, 'два'),
 Substring(162, 171, 'вертолёта')]
```

```
[_.text for _ in n_tok_text]
```

```
['Сикорский',
 'родился',
```

```
 '7',
 'июня',
 '1889',
 'года',
 '.',
 'Он',
 'поступил',
 'в',
 'Киевский',
 'политехнический',
 'институт',
 'в',
 '1907',
 'году',
 '.',
 'В',
 '1909-1912',
 'годах',
 'студент',
 'Сикорский',
 'спроектировал',
 'и',
 'построил',
 'два',
 'вертолёта']
```

```
n_sen_text = list(sentenize(text))
n_sen_text
```

```
[Substring(0, 35, 'Сикорский родился 7 июня 1889 года.'),
 Substring(36,
           96,
           'Он поступил в Киевский политехнический инст
 Substring(97,
           171,
           'В 1909-1912 годах студент Сикорский спроектиров
```

```
[_.text for _ in n_sen_text], len([_.text for _ in n_sen_text])
```

```
(['Сикорский родился 7 июня 1889 года.',
  'Он поступил в Киевский политехнический институт
  'В 1909-1912 годах студент Сикорский спроектировал и п
 3)
```

```
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
```

```
n_sen_chunk = n_sentenize(text)
n_sen_chunk
```

```
[['С и к о р с к и й', 'р о д и л с я', '7', 'и ю н я', '1889', 'г о д а', '.'],
 ['О н',
  'п о с т у п и л',
  'в',
  'К и е в с к и й',
  'п о л и т е х н и ч е с к и й',
  'и н с т и т у т',
  'в',
  '1907',
  'г о д у',
  '.'],
 ['В',
  '1909-1912',
  'г о д а х',
  'с т у д е н т',
  'С и к о р с к и й',
  'с п р о е к т и р о в а л',
  'и',
  'п о с т р о и л',
  'д в а',
  'в е р т о л ё т а']]
```

Частеречная разметка

```
!pip  install  navec
!pip  install  slovnet
from  navec  import  Navec
from  slovnet  import  Morph
```

```
Requirement already satisfied: navec in /usr/local/lib/python3.7/dist-packages (0.10.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from navec) (
Collecting slovnet
  Downloading https://files.pythonhosted.org/packages/a9/3b/f1ef495be8990004959dd0510c95f688d
     |████████████████████████████████| 51kB 1.6MB/s
Requirement already satisfied: navec in /usr/local/lib/python3.7/dist-packages (from slovnet)
Requirement already satisfied: razdel in /usr/local/lib/python3.7/dist-packages (from slovnet
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from slovnet)
Installing collected packages: slovnet
Successfully installed slovnet-0.5.0
```

```
%cd  /content/drive/MyDrive
```

```
/content/drive/MyDrive
```

```
navec  =  Navec.load('navec_news_v1_1B_250K_300d_100q.tar')
```

```
n_morph  =  Morph.load('slovnet_morph_news_v1.tar',  batch_size=4)
```

```
morph_res  =  n_morph.navec(navec)
```

```
def  print_pos(markup):
```

```
        for token in markup.tokens:
            print('{} - {}'.format(token.text, token.tag))


n_text_markup = list(_ for _ in n_morph.map(n_sen_chunk))
[print_pos(x) for x in n_text_markup]
```

```
С и к о р с к и й - PROPN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
р о д и л с я - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Vc
7 - ADJ
и ю н я - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
1889 - ADJ
г о д а - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
. - PUNCT
О н - PRON|Case=Nom|Gender=Masc|Number=Sing|Person=3
п о с т у п и л - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|
в - ADP
К и е в с к и й - ADJ|Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing
п о л и т е х н и ч е с к и й - ADJ|Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing
и н с т и т у т - NOUN|Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing
в - ADP
1907 - ADJ
г о д у - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
. - PUNCT
В - ADP
1909-1912 - ADJ
г о д а х - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Plur
с т у д е н т - NOUN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
С и к о р с к и й - PROPN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
с п р о е к т и р о в а л - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|Ver
и - CCONJ
п о с т р о и л - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|
д в а - NUM|Animacy=Inan|Case=Acc|Gender=Masc
в е р т о л ё т а - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
[None, None, None]
```

Лемматизация

```
!pip install natasha
from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, MorphVocab
```

```
Collecting natasha
  Downloading https://files.pythonhosted.org/packages/51/8e/ab0745100be276750fb6b8858c6180a17
     |████████████████████████████████| 34.4MB 1.5MB/s
Requirement already satisfied: razdel>=0.5.0 in /usr/local/lib/python3.7/dist-packages (from
Collecting yargy>=0.14.0
  Downloading https://files.pythonhosted.org/packages/d3/46/bc1a17200a55f4b0608f39ac64f1840fd
     |████████████████████████████████| 51kB 5.5MB/s
Collecting pymorphy2
  Downloading https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9886b64a54b4
     |████████████████████████████████| 61kB 7.0MB/s
Collecting ipymarkup>=0.8.0
  Downloading https://files.pythonhosted.org/packages/bf/9b/bf54c98d50735a4a7c84c71e92c536173
Requirement already satisfied: slovnet>=0.3.0 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: navec>=0.9.0 in /usr/local/lib/python3.7/dist-packages (from n
```

```
Collecting dawg-python>=0.7.1
  Downloading https://files.pythonhosted.org/packages/6a/84/ff1ce2071d4c650ec85745766c0047ccc
Collecting pymorphy2-dicts-ru<3.0,>=2.4
  Downloading https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9e96badf174
     |████████████████████████████████| 8.2MB 16.1MB/s
Requirement already satisfied: docopt>=0.6 in /usr/local/lib/python3.7/dist-packages (from py
Collecting intervaltree>=3
  Downloading https://files.pythonhosted.org/packages/50/fb/396d568039d21344639db96d940d40eb
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from slovnet>
Requirement already satisfied: sortedcontainers<3.0,>=2.0 in /usr/local/lib/python3.7/dist-pa
Building wheels for collected packages: intervaltree
  Building wheel for intervaltree (setup.py) ... done
  Created wheel for intervaltree: filename=intervaltree-3.1.0-py2.py3-none-any.whl size=26102
  Stored in directory: /root/.cache/pip/wheels/f3/f2/66/e9c30d3e9499e65ea2fa0d07c002e64de63bd
Successfully built intervaltree
Installing collected packages: dawg-python, pymorphy2-dicts-ru, pymorphy2, yargy, intervaltre
  Found existing installation: intervaltree 2.1.0
    Uninstalling intervaltree-2.1.0:
      Successfully uninstalled intervaltree-2.1.0
Successfully installed dawg-python-0.7.2 intervaltree-3.1.0 ipymarkup-0.9.0 natasha-1.4.0 pym
```

```
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


n_doc = n_lemmatize(text)
{_.text: _.lemma for _ in n_doc.tokens}
```

```
{'.': '.',
 '1889': '1889',
 '1907': '1907',
 '1909-1912': '1909-1912',
 '7': '7',
 'В': 'в',
 'Киевский': 'киевский',
 'Он': 'он',
 'Сикорский': 'сикорский',
 'в': 'в',
 'вертолёта': 'вертолет',
 'года': 'год',
 'годах': 'год',
 'году': 'год',
 'два': 'два',
 'и': 'и',
 'институт': 'институт',
 'июня': 'июнь',
 'политехнический': 'политехнический',
 'построил': 'построить',
```

```
 'поступил': 'поступить',
 'родился': 'родиться',
 'спроектировал': 'спроектировать',
 'студент': 'студент'}
```

Выделение (распознавание) именованных сущностей, named-entity recognition (NER)

```
from slovnet import NER
from ipymarkup import show_span_ascii_markup as show_markup

ner = NER.load('slovnet_ner_news_v1.tar')

ner_res = ner.navec(navec)

markup_ner = ner(text)
markup_ner
```

```
SpanMarkup(
    text='Сикорский родился 7 июня 1889 года. Он поступил в
    spans=[Span(
         start=50,
         stop=83,
         type='ORG'
     ), Span(
         start=123,
         stop=132,
         type='PER'
     )]
)
```

```
show_markup(markup_ner.text, markup_ner.spans)
```

```
Сикорский родился 7 июня 1889 года. Он поступил в Киев
                                    ORG──────
политехнический институт в 1907 году. В 1909-1912 годах с
──────────────────────
Сикорский спроектировал и построил два вертолёта
PER──────
```

Разбор предложения

```
from natasha import NewsSyntaxParser

emb = NewsEmbedding()
```

```
emb = NewsEmbedding()
syntax_parser = NewsSyntaxParser(emb)
```

```
n_doc.parse_syntax(syntax_parser)
n_doc.sents[0].syntax.print()
```

```
          ┌► Сикорский nsubj
┌──────── ┌── └── родился
│  ┌── └► ┌── 7         obl
│  │      └► июня       flat
│  │      ┌► 1889       amod
│  └────► └── года      nmod
└────────────► .        punct
```

```
n_doc.sents[1].syntax.print()
```

```
            ┌► Он                     nsubj
┌── ┌── ┌──────────── └── поступил
│   │   │  ┌─────────► в                   case
│   │   │  │  ┌─────► Киевский              amod
│   │   │  │  │  ┌► политехнический amod
│   │   └► └── └── └── институт          obl
│   │       ┌─────► в                    case
│   │       │  ┌► 1907                  amod
│   └──────────► └── └── году                 obl
└──────────────────────► .                    punct
```

```
n_doc.sents[2].syntax.print()
```

```
     ┌─────► В                case
     │  ┌► 1909-1912      amod
  ┌► └── └── годах          obl
  │  ┌─────► студент        nsubj
  │  │  ┌► Сикорский      nsubj
  └── └── └── спроектировал
  │     ┌► и                cc
┌── └─────► └── построил       conj
│       ┌► два            nummod:gov
└──────────► └── вертолёта     obj
```