**ЛР№5 Предобработки текста**

ИУ5И-21М  
Для выполнения работы испоьзована библиотека 'Natasha'

[1]

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'

**Задача токенизации**

[4]

!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/664223a3924aa6e70479f7d37220b3a658765b9cfe760b4af7ffdc50d38f/razdel-0.5.0-py3-none-any.whl  
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, 'годах'),  
 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')]

[6]

[\_.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']

[7]

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 институт в 1907 году.'),  
 Substring(97,  
 171,  
 'В 1909-1912 годах студент Сикорский спроектировал и построил два вертолёта')]

[8]

[\_.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 1907 году.',  
 'В 1909-1912 годах студент Сикорский спроектировал и построил два вертолёта'],  
 3)

[9]

**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

[10]

n\_sen\_chunk = n\_sentenize(text)  
n\_sen\_chunk

[['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']]

**Частеречная разметка**

[13]

!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) (1.19.5)  
Collecting slovnet  
?25l Downloading https://files.pythonhosted.org/packages/a9/3b/f1ef495be8990004959dd0510c95f688d1b07529f6a862bc56a405770b26/slovnet-0.5.0-py3-none-any.whl (49kB)  
K |████████████████████████████████| 51kB 1.6MB/s   
?25hRequirement already satisfied: navec in /usr/local/lib/python3.7/dist-packages (from slovnet) (0.10.0)  
Requirement already satisfied: razdel in /usr/local/lib/python3.7/dist-packages (from slovnet) (0.5.0)  
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from slovnet) (1.19.5)  
Installing collected packages: slovnet  
Successfully installed slovnet-0.5.0

[21]

%cd /content/drive/MyDrive

/content/drive/MyDrive

[22]

navec = Navec.load('navec\_news\_v1\_1B\_250K\_300d\_100q.tar')

[23]

n\_morph = Morph.load('slovnet\_morph\_news\_v1.tar', batch\_size=4)

[24]

morph\_res = n\_morph.navec(navec)

[25]

**def** print\_pos(markup):  
    **for** token **in** markup.tokens:  
        print('{} - {}'.format(token.text, token.tag))

[28]

n\_text\_markup = list(\_ **for** \_ **in** n\_morph.map(n\_sen\_chunk))  
[print\_pos(x) **for** x **in** n\_text\_markup]

Сикорский - PROPN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing  
родился - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Mid  
7 - ADJ  
июня - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing  
1889 - ADJ  
года - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing  
. - PUNCT  
Он - PRON|Case=Nom|Gender=Masc|Number=Sing|Person=3  
поступил - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  
в - ADP  
Киевский - ADJ|Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing  
политехнический - ADJ|Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing  
институт - NOUN|Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing  
в - ADP  
1907 - ADJ  
году - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing  
. - PUNCT  
В - ADP  
1909-1912 - ADJ  
годах - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Plur  
студент - NOUN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing  
Сикорский - PROPN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing  
спроектировал - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  
и - CCONJ  
построил - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  
два - NUM|Animacy=Inan|Case=Acc|Gender=Masc  
вертолёта - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing

[None, None, None]

**Лемматизация**

[31]

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

Collecting natasha  
?25l Downloading https://files.pythonhosted.org/packages/51/8e/ab0745100be276750fb6b8858c6180a1756696572295a74eb5aea77f3bbd/natasha-1.4.0-py3-none-any.whl (34.4MB)  
K |████████████████████████████████| 34.4MB 1.5MB/s   
?25hRequirement already satisfied: razdel>=0.5.0 in /usr/local/lib/python3.7/dist-packages (from natasha) (0.5.0)  
Collecting yargy>=0.14.0  
?25l Downloading https://files.pythonhosted.org/packages/d3/46/bc1a17200a55f4b0608f39ac64f1840fd4a52f9eeea462d9afecbf71246b/yargy-0.15.0-py3-none-any.whl (41kB)  
K |████████████████████████████████| 51kB 5.5MB/s   
?25hCollecting pymorphy2  
?25l Downloading https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9886b64a54b476e1a332c67eee9f88e7f1ae8c9/pymorphy2-0.9.1-py3-none-any.whl (55kB)  
K |████████████████████████████████| 61kB 7.0MB/s   
?25hCollecting ipymarkup>=0.8.0  
 Downloading https://files.pythonhosted.org/packages/bf/9b/bf54c98d50735a4a7c84c71e92c5361730c878ebfe903d2c2d196ef66055/ipymarkup-0.9.0-py3-none-any.whl  
Requirement already satisfied: slovnet>=0.3.0 in /usr/local/lib/python3.7/dist-packages (from natasha) (0.5.0)  
Requirement already satisfied: navec>=0.9.0 in /usr/local/lib/python3.7/dist-packages (from natasha) (0.10.0)  
Collecting dawg-python>=0.7.1  
 Downloading https://files.pythonhosted.org/packages/6a/84/ff1ce2071d4c650ec85745766c0047ccc3b5036f1d03559fd46bb38b5eeb/DAWG\_Python-0.7.2-py2.py3-none-any.whl  
Collecting pymorphy2-dicts-ru<3.0,>=2.4  
?25l Downloading https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9e96badf174068a2dd20264b9a378f2be1cdd9e/pymorphy2\_dicts\_ru-2.4.417127.4579844-py2.py3-none-any.whl (8.2MB)  
K |████████████████████████████████| 8.2MB 16.1MB/s   
?25hRequirement already satisfied: docopt>=0.6 in /usr/local/lib/python3.7/dist-packages (from pymorphy2->natasha) (0.6.2)  
Collecting intervaltree>=3  
 Downloading https://files.pythonhosted.org/packages/50/fb/396d568039d21344639db96d940d40eb62befe704ef849b27949ded5c3bb/intervaltree-3.1.0.tar.gz  
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from slovnet>=0.3.0->natasha) (1.19.5)  
Requirement already satisfied: sortedcontainers<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from intervaltree>=3->ipymarkup>=0.8.0->natasha) (2.3.0)  
Building wheels for collected packages: intervaltree  
 Building wheel for intervaltree (setup.py) ... ?25l?25hdone  
 Created wheel for intervaltree: filename=intervaltree-3.1.0-py2.py3-none-any.whl size=26102 sha256=2ebf2fe745648b6116032daa362816912c6b841ba054efbe4a3337d47855acaa  
 Stored in directory: /root/.cache/pip/wheels/f3/f2/66/e9c30d3e9499e65ea2fa0d07c002e64de63bd0adaa49c445bf  
Successfully built intervaltree  
Installing collected packages: dawg-python, pymorphy2-dicts-ru, pymorphy2, yargy, intervaltree, ipymarkup, natasha  
 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 pymorphy2-0.9.1 pymorphy2-dicts-ru-2.4.417127.4579844 yargy-0.15.0

[32]

**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

[35]

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)**

[36]

**from** slovnet **import** NER  
**from** ipymarkup **import** show\_span\_ascii\_markup **as** show\_markup

[38]

ner = NER.load('slovnet\_ner\_news\_v1.tar')

[39]

ner\_res = ner.navec(navec)

[41]

markup\_ner = ner(text)  
markup\_ner

SpanMarkup(  
 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',  
 spans=[Span(  
 start=50,  
 stop=83,  
 type='ORG'  
 ), Span(  
 start=123,  
 stop=132,  
 type='PER'  
 )]  
)

[42]

show\_markup(markup\_ner.text, markup\_ner.spans)

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

**Разбор предложения**

[43]

**from** natasha **import** NewsSyntaxParser

[44]

emb = NewsEmbedding()  
syntax\_parser = NewsSyntaxParser(emb)

[45]

n\_doc.parse\_syntax(syntax\_parser)  
n\_doc.sents[0].syntax.print()

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

[46]

n\_doc.sents[1].syntax.print()

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

[47]

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