Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №5
по дисциплине
«Методы машинного обучения»

Выполнил:
студент группы ИУ5-22М
ЧжаоЛян

Москва — 2022 г.

Задача токенизации

```
In [1]: text='В 1876 году Чехов-старший разорился, и вся семья уехала в Мо
```

Задача токенизации

```
In [2]: !pip install razdel
        from razdel import tokenize, sentenize
        n_tok_text = list(tokenize(text))
        n_tok_text
```

```
Requirement already satisfied: razdel in c:\zl\work\anaconda\anaconda\lib\site-packages (0.5.0)

Out[2]: [Substring(0, 1, 'В'),
         Substring(2, 6, '1876'),
         Substring(7, 11, 'году'),
         Substring(12, 25, 'Чехов-старший'),
         Substring(26, 35, 'разорился'),
         Substring(35, 36, ','),
         Substring(37, 38, 'и'),
         Substring(39, 42, 'вся'),
         Substring(43, 48, 'семья'),
         Substring(49, 55, 'уехала'),
         Substring(56, 57, 'в'),
         Substring(58, 64, 'Москву'),
         Substring(64, 65, '.'),
         Substring(66, 83, 'Шестнадцатилетний'),
         Substring(84, 89, 'Антон'),
         Substring(89, 90, ','),
         Substring(91, 102, 'завершавший'),
         Substring(103, 111, 'обучение')
```

```
In [3]: [_.text for _ in n_tok_text]
```

```
'годы',
'он',
'много',
'читал',
',',
'писал',
'очерки',
'для',
'гимназических',
'журналов',
',',
'а',
'журнал',
'«',
'Заика',
'»',
'с',
'короткими',
```

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

Out[4]:
```
[Substring(0,
           65,
           'В 1876 году Чехов-старший разорился, и вся семья уехала в Москву.'),
 Substring(66,
           196,
           'Шестнадцатилетний Антон, завершавший обучение в гимназии, остался один и занимался репетиторством, чтобы заработать себе на жизнь.'),
 Substring(197,
           355,
           'В эти годы он много читал, писал очерки для гимназических журналов, а журнал «Заика» с короткими зарисовками из таганрогской жизни отправлял братьям в Москву.'),
 Substring(356,
           507,
           'Тогда же Чехов написал первую пьесу — «Безотцовщина» и водевиль «Недаром курица пела». В 1879 году Чехов окончил гимназию и уехал из Таганрога в Москву.'),
 Substring(508,
           606,
           'Там он начал заботиться о семье, обеспечивал близких на скромный доход от литературных публикаций.'),
 Substring(607,
           904,
           'Дебют Чехова в печати состоялся в декабре того же года: в журнале «Стрекоза» были опубликованы рассказ «Письмо к ученому соседу» и юмореска «Что чаще всего встречается в романах, повестях и т. п.». В этом же году Чехов поступил на медицинский факультет Московского университета имени И. М. Сеченова.'),
 Substring(905,
           987,
           'Студент-медик жил у брата Ивана в подмосковном Воскресенске (сегодня город Истра).'),
 Substring(988,
           1099,
           'Там же в 1881 году он познакомился с заведующим Воскресенской земской больницей, доктором Павлом Архангельским.'),
 Substring(1100,
           1232,
           'Еще во время учебы Чехов принимал больных, здесь же прошел практику, а после окончания университета остался работать уездным врачом.'),
 Substring(1233,
           1310,
           'Летом 1884 года он перешел на должность заведую
```

In [5]:
```
[_.text for _ in n_sen_text], len([_.text for _ in n_sen_text])
```

Out[5]:
```
(['В 1876 году Чехов-старший разорился, и вся семья уехала в Москву.',
  'Шестнадцатилетний Антон, завершавший обучение в гимназии, остался один и занимался репетиторством, чтобы заработать себе на жизнь.',
  'В эти годы он много читал, писал очерки для гимназических журналов, а журнал «Заика» с короткими зарисовками из таганрогской жизни отправлял братьям в Москву.',
  'Тогда же Чехов написал первую пьесу — «Безотцовщина» и водевиль «Недаром курица пела». В 1879 году Чехов окончил гимназию и уехал из Таганрога в Москву.',
  'Там он начал заботиться о семье, обеспечивал близких на скромный доход от литературных публикаций.',
  'Дебют Чехова в печати состоялся в декабре того же года: в журнале «Стрекоза» были опубликованы рассказ «Письмо к ученому соседу» и юмореска «Что чаще всего встречается в романах, повестях и т. п.». В этом же году Чехов поступил на медицинский факультет Московского университета имени И. М. Сеченова.',
  'Студент-медик жил у брата Ивана в подмосковном Воскресенске (сегодня город Истра).',
  'Там же в 1881 году он познакомился с заведующим Воскресенской земской больницей, доктором Павлом Архангельским.',
  'Еще во время учебы Чехов принимал больных, здесь же прошел практику, а после окончания университета остался работать уездным врачом.',
  'Летом 1884 года он перешел на должность заведующего звенигородской больницей.'],
 10)
```

In [6]:
```
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
```

In [7]:
```
n_sen_chunk = n_sentenize(text)
n_sen_chunk
```

```
 'и',
 'занимался',
 'репетиторством',
 ',',
 'чтобы',
 'заработать',
 'себе',
 'на',
 'жизнь',
 '.'],
['В',
 'эти',
 'годы',
 'он',
 'много',
 'читал',
 ',',
 'писал',
 'очерки',
 'для',
```

Частеречная разметка

In [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
  Downloading https://files.pythonhosted.org/packages/a9/3b/f1ef495be8990004959dd0510c95f688d1b07529f6a862bc56a405770b26/slovnet-0.5.0-py3-none-any.whl (49kB)
     |████████████████████████████████| 51kB 1.6MB/s
Requirement 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
```

In [21]:
```
%cd /content/drive/MyDrive
```

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

In [22]:
```
navec = Navec.load('navec_news_v1_1B_250K_300d_100q.tar')
```

In [23]:
```
n_morph = Morph.load('slovnet_morph_news_v1.tar', batch_size=4)
```

In [24]:
```
morph_res = n_morph.navec(navec)
```

In [25]:
```
def print_pos(markup):
    for token in markup.tokens:
        print('{} - {}'.format(token.text, token.tag))
```

In [28]:
```
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|Voice=Mid
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|Voice=Act
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|VerbForm=Fin|Voice=Act
и - CCONJ
п о с т р о и л - VERB|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act
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
```

Out[28]:
```
[None, None, None]
```

Лемматизация

In [31]:
```
!pip install natasha
from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, MorphVocab
```

```
Collecting natasha
  Downloading https://files.pythonhosted.org/packages/51/8e/ab0745100be276750fb6b8858c6180a1756696572295a74eb5aea77f3bbd/natasha-1.4.0-py3-none-any.whl (34.4MB)
     |████████████████████████████████| 34.4MB 1.5MB/s
Requirement already satisfied: razdel>=0.5.0 in /usr/local/lib/python3.7/dist-packages (from natasha) (0.5.0)
Collecting yargy>=0.14.0
  Downloading https://files.pythonhosted.org/packages/d3/46/bc1a17200a55f4b0608f39ac64f1840fd4a52f9eeea462d9afecbf71246b/yargy-0.15.0-py3-none-any.whl (41kB)
     |████████████████████████████████| 51kB 5.5MB/s
Collecting pymorphy2
  Downloading https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9886b64a54b476e1a332c67eee9f88e7f1ae8c9/pymorphy2-0.9.1-py3-none-any.whl (55kB)
     |████████████████████████████████| 61kB 7.0MB/s
Collecting 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
  Downloading https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9e96badf174068a2dd20264b9a378f2be1cdd9e/pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl (8.2MB)
     |████████████████████████████████| 8.2MB 16.1MB/s
Requirement 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) ... done
  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
```

In [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
```

In [35]:
```
n_doc = n_lemmatize(text)
{_.text: _.lemma for _ in n_doc.tokens}
```

Out[35]:
```
{'.': '.',
 '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)

```
In [36]: from slovnet import NER
         from ipymarkup import show_span_ascii_markup as show_markup

In [38]: ner = NER.load('slovnet_ner_news_v1.tar')

In [39]: ner_res = ner.navec(navec)

In [41]: markup_ner = ner(text)
         markup_ner

Out[41]: SpanMarkup(
             text='Сикорский родился 7 июня 1889 года. Он поступил в Киевски
         й политехнический институт в 1907 году. В 1909-1912 годах студент С
         икорский спроектировал и построил два вертолёта',
             spans=[Span(
                      start=50,
                      stop=83,
                      type='ORG'
                  ), Span(
                      start=123,
                      stop=132,
                      type='PER'
                  )]
         )

In [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

In [43]:
```
from natasha import NewsSyntaxParser
```

In [44]:
```
emb = NewsEmbedding()
syntax_parser = NewsSyntaxParser(emb)
```

In [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
```

In [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
```

In [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
```

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Подготовка обучающей и тестовой выборки,
кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»
[Электронный ресурс] // GitHub. — 2019. — Режим доступа:
https://github.com/
ugapanyuk/ml_course/wiki/LAB_KNN (дата обращения: 05.04.2019).
[2] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource] //
Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/
stable/ (online; accessed: 20.02.2019).

[3] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. —

Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).
[4] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode:
http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).
[5] dronio. Solar Radiation Prediction [Electronic resource] // Kaggle. — 2017. — Access
mode: https://www.kaggle.com/dronio/SolarEnergy (online; accessed: 18.02.2019).
[6] Chrétien M. Convert datetime.time to seconds [Electronic resource] // Stack Overflow.
— 2017. — Access mode: https://stackoverflow.com/a/44823381 (online; accessed:
20.02.2019).
[7] scikit-learn 0.20.3 documentation [Electronic resource]. — 2019. — Access mode: https:
//scikit-learn.org/ (online; accessed: 05.04.2019).