

Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



Лабораторная работа № 8

по дисциплине «Методы машинного обучения»

Предобработка текста

ИСПОЛНИТЕЛЬ:

студент ИУ5-23М

Бондаренко И. Г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю. Е.

___ " _____ " 2024 г.

Москва, 2024

Задание лабораторной работы

Для произвольного предложения или текста решить следующие задачи:

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

Выполнение работы

Исходный текст:

In [34]:

```
text = 'Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе. Т
```

Токенизация

In [35]:

```
!pip install nltk
```

Requirement already satisfied: nltk in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (3.8.1)

Requirement already satisfied: click in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from nltk) (8.1.7)

Requirement already satisfied: joblib in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from nltk) (1.4.2)

Requirement already satisfied: regex<=2021.8.3 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from nltk) (2024.5.15)

Requirement already satisfied: tqdm in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from nltk) (4.66.4)

Requirement already satisfied: colorama in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from click->nltk) (0.4.6)

In [36]:

```
import nltk
```

```
from nltk.tokenize import Punkt
```

```
nltk.download('punkt')
```

[nltk_data] Downloading package punkt to

[nltk_data] C:\Users\firry\AppData\Roaming\nltk_data...

[nltk_data] Package punkt is already up-to-date!

Out[36]:

True

In [37]:

```
from nltk import tokenize
```

```
dir(tokenize)[:18]
```

Out[37]:

['BlanklineTokenizer',

'LegalitySyllableTokenizer',

'LineTokenizer',

'MWETokenizer',

'NLTKWordTokenizer',

'PunktSentenceTokenizer',

'RegexTokenizer',

'ReppTokenizer',

'SEXPRTokenizer',

'SpaceTokenizer',

'StanfordSegmenter',

'SyllableTokenizer',

'TabTokenizer',

'TextTilingTokenizer',

'ToktokTokenizer',

'TreebankWordDetokenizer',

'TreebankWordTokenizer',

'TweetTokenizer']

In [38]:

```
nltk_tokenize = nltk.WordPunctTokenizer()
```

```
nltk_tokenize.tokenize(text)
```

```

Out[38]:
['Пятнадцатизэтажный',
'музейный',
'комплекс',
'с',
'концертным',
'залом',
'построят',
'на',
'Малой',
'Почтовой',
'улице',
'в',
'Басманном',
'районе',
'',
'Там',
'разместят',
'частную',
'коллекцию',
'произведений',
'искусства',
'',
'4',
'этажа',
'займут',
'офисы',
'',
'а',
'верхний',
'уровень',
'отведен',
'под',
'ресторан',
'с',
'видовой',
'террасой',
'.']

```

Токенизация по предложениям:

```

In [39]:
nltk_tokenize.sent_tokenize(text)
print(len(nltk_tokenize.sent_tokenize(text)))
nltk_tokenize.sent_tokenize(text)

```

3

```

Out[39]:
['Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе.',
'Там разместят частную коллекцию произведений искусства.',
'4 этажа займут офисы, а верхний уровень отведен под ресторан с видовой террасой.']

```

```

In [40]:

```

```

!pip install razdel

```

```

Requirement already satisfied: razdel in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (0.5.0)

```

```

In [41]:

```

```

from razdel import tokenize, sentenize

```

```

In [42]:

```

```

n_tok_text = list(tokenize(text))

```

```

n_tok_text

```

Out[42]:

```
[Substring(0, 17, 'Пятинадцатизэтажный'),  
Substring(18, 26, 'музейный'),  
Substring(27, 35, 'комплекс'),  
Substring(36, 37, 'с'),  
Substring(38, 48, 'концертным'),  
Substring(49, 54, 'залом'),  
Substring(55, 63, 'построят'),  
Substring(64, 66, 'на'),  
Substring(67, 72, 'Малой'),  
Substring(73, 81, 'Почтовой'),  
Substring(82, 87, 'улице'),  
Substring(88, 89, 'в'),  
Substring(90, 99, 'Басманном'),  
Substring(100, 106, 'районе'),  
Substring(106, 107, '.'),  
Substring(108, 111, 'Там'),  
Substring(112, 121, 'разместят'),  
Substring(122, 129, 'частную'),  
Substring(130, 139, 'коллекцию'),  
Substring(140, 152, 'произведений'),  
Substring(153, 162, 'искусства'),  
Substring(162, 163, '.'),  
Substring(164, 165, '4'),  
Substring(166, 171, 'этажа'),  
Substring(172, 178, 'займут'),  
Substring(179, 184, 'офисы'),  
Substring(184, 185, ','),  
Substring(186, 187, 'а'),  
Substring(188, 195, 'верхний'),  
Substring(196, 203, 'уровень'),  
Substring(204, 211, 'отведен'),  
Substring(212, 215, 'под'),  
Substring(216, 224, 'ресторан'),  
Substring(225, 226, 'с'),  
Substring(227, 234, 'видовой'),  
Substring(235, 243, 'террасой'),  
Substring(243, 244, '.')] ]
```

In [43]:

```
[_text for _ in n_tok_text]
```

Out[43]:

```
['Пятинадцатизэтажный',  
'музейный',  
'комплекс',  
'с',  
'концертным',  
'залом',  
'построят',  
'на',  
'Малой',  
'Почтовой',  
'улице',  
'в',  
'Басманном',  
'районе',  
'.',  
'Там',  
'разместят',  
'частную',  
'коллекцию',  
'произведений',  
'искусства',  
'',  
'4',  
'этажа',  
'займут',  
'офисы',  
'',  
'а',  
'верхний',  
'уровень',  
'отведен',  
'под',  
'ресторан',  
'с',  
'видовой',  
'террасой',  
''] ]
```

In [44]:

```
n_sen_text = list(sentenize(text))
n_sen_text
Out[44]:
[Substring(0,
  107,
  'Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе.'),
Substring(108,
  163,
  'Там разместят частную коллекцию произведений искусства.'),
Substring(164,
  244,
  '4 этажа займут офисы, а верхний уровень отведен под ресторан с видовой террасой.')]

```

```
In [45]:
[_.text for _ in n_sen_text], len([_.text for _ in n_sen_text])
Out[45]:
(['Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе.',
 'Там разместят частную коллекцию произведений искусства.',
 '4 этажа займут офисы, а верхний уровень отведен под ресторан с видовой террасой.'],
3)

```

Токенизация для последующей обработки:

```
In [46]:
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 [47]:
n_sen_chunk = n_sentenize(text)
n_sen_chunk

```

```
Out[47]:
[['Пятнадцатизэтажный',
 'музейный',
 'комплекс',
 'с',
 'концертным',
 'залом',
 'построят',
 'на',
 'Малой',
 'Почтовой',
 'улице',
 'в',
 'Басманном',
 'районе',
 ''],
 ['Там',
 'разместят',
 'частную',
 'коллекцию',
 'произведений',
 'искусства',
 ''],
 ['4',
 'этажа',
 'займут',
 'офисы',
 ',',
 'а',
 'верхний',
 'уровень',
 'отведен',
 'под',
 'ресторан',
 'с',
 'видовой',
 'террасой',
 '']]

```

Частичная разметка

```
In [48]:
!pip install navec
!pip install slovnet

```

Requirement already satisfied: navec in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (0.10.0)
Requirement already satisfied: numpy in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from navec) (1.26.4)
Requirement already satisfied: slovnet in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (0.6.0)
Requirement already satisfied: numpy in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from slovnet) (1.26.4)
Requirement already satisfied: razdel in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from slovnet) (0.5.0)
Requirement already satisfied: navec in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from slovnet) (0.10.0)

In [49]:

```
from navec import Navec
```

```
from slovnet import Morph
```

In [50]:

```
# Файл необходимо скачать по ссылке https://github.com/natasha/navec#downloads
```

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

In [51]:

```
# Файл необходимо скачать по ссылке https://github.com/natasha/slovnet#downloads
```

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

In [52]:

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

In [53]:

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

```
    for token in markup.tokens:
```

```
        print('{} - {}'.format(token.text, token.tag))
```

In [54]:

```
n_text_markup = list(_for_ in n_morph.map(n_sen_chunk))
```

```
[print_pos(x) for x in n_text_markup]
```

Пятнадцатизатяжной - ADJ|Case=Nom|Degree=Pos|Gender=Masc|Number=Sing
музейный - ADJ|Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing
комплекс - NOUN|Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing
с - ADP
концертным - ADJ|Case=Ins|Degree=Pos|Gender=Masc|Number=Sing
залом - NOUN|Animacy=Inan|Case=Ins|Gender=Masc|Number=Sing
построят - VERB|Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin|Voice=Act
на - ADP
Малой - ADJ|Case=Loc|Degree=Pos|Gender=Fem|Number=Sing
Почтовой - ADJ|Case=Loc|Degree=Pos|Gender=Fem|Number=Sing
улице - NOUN|Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing
в - ADP
Басманном - ADJ|Case=Loc|Degree=Pos|Gender=Masc|Number=Sing
районе - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
. - PUNCT
Там - ADV|Degree=Pos
разместят - VERB|Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin|Voice=Act
частную - ADJ|Case=Acc|Degree=Pos|Gender=Fem|Number=Sing
коллекцию - NOUN|Animacy=Inan|Case=Acc|Gender=Fem|Number=Sing
произведений - NOUN|Animacy=Inan|Case=Gen|Gender=Neut|Number=Plur
искусства - NOUN|Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing
. - PUNCT
4 - NUM
этажа - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
займут - VERB|Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin|Voice=Act
офисы - NOUN|Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur
, - PUNCT
а - CCONJ
верхний - ADJ|Case=Nom|Degree=Pos|Gender=Masc|Number=Sing
уровень - NOUN|Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing
отведен - VERB|Aspect=Perf|Gender=Masc|Number=Sing|Tense=Past|Variant=Short|VerbForm=Part|Voice=Pass
под - ADP
ресторан - NOUN|Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing
с - ADP
видовой - ADJ|Case=Ins|Degree=Pos|Gender=Fem|Number=Sing
террасой - NOUN|Animacy=Inan|Case=Ins|Gender=Fem|Number=Sing
. - PUNCT
Out[54]:
[None, None, None]

Лемматизация

In [55]:

```
!pip install natasha
```

```
!pip install setuptools
```

Requirement already satisfied: natasha in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (1.6.0)
 Requirement already satisfied: pymorphy2 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.9.1)
 Requirement already satisfied: razdel>=0.5.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.5.0)
 Requirement already satisfied: navec>=0.9.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.10.0)
 Requirement already satisfied: slovnets>=0.6.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.6.0)
 Requirement already satisfied: yargy>=0.16.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.16.0)
 Requirement already satisfied: ipymarkup>=0.8.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from natasha) (0.9.0)
 Requirement already satisfied: intervaltree>=3 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from ipymarkup>=0.8.0->natasha) (3.1.0)
 Requirement already satisfied: numpy in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from navec>=0.9.0->natasha) (1.26.4)
 Requirement already satisfied: dawg-python>=0.7.1 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from pymorphy2->natasha) (0.7.2)
 Requirement already satisfied: pymorphy2-dicts-ru<3.0,>=2.4 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from pymorphy2->natasha) (2.4.417127.4579844)
 Requirement already satisfied: docopt>=0.6 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from pymorphy2->natasha) (0.6.2)
 Requirement already satisfied: sortedcontainers<3.0,>=2.0 in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (from intervaltree>=3->ipymarkup>=0.8.0->natasha) (2.4.0)
 Requirement already satisfied: setuptools in c:\users\firry\appdata\local\programs\python\python312\lib\site-packages (70.0.0)

In [56]:

from natasha import Doc, Segmenter, NewsEmbedding, NewsMorphTagger, MorphVocab

In [57]:

```
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 [58]:

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

Out[58]:

```
{'Пятнадцатизэтажный': 'пятнадцатизэтажный',
'музейный': 'музейный',
'комплекс': 'комплекс',
'с': 'с',
'концертным': 'концертный',
'залом': 'зал',
'построят': 'построить',
'на': 'на',
'Малой': 'малый',
'Почтовой': 'почтовый',
'улице': 'улица',
'в': 'в',
'Басманном': 'басманный',
'районе': 'район',
'': ''',
'Там': 'там',
'разместят': 'разместить',
'частную': 'частный',
'коллекцию': 'коллекция',
'произведений': 'произведение',
'искусства': 'искусство',
'4': '4',
'этажа': 'этаж',
'займут': 'занять',
'офисы': 'офис',
'': ''',
'а': 'а',
'верхний': 'верхний',
'уровень': 'уровень',
'отведен': 'отвести',
'под': 'под',
'ресторан': 'ресторан',
'видовой': 'видовой',
'террасой': 'терраса'}
```

Выделение (распознавание) именованных сущностей

In [59]:

!pip install ipymarkup

Requirement already satisfied: ipymarkup in c:\users\firry\AppData\Local\Programs\Python\Python312\lib\site-packages (0.9.0)
Requirement already satisfied: intervaltree>=3 in c:\users\firry\AppData\Local\Programs\Python\Python312\lib\site-packages (from ipymarkup) (3.1.0)
Requirement already satisfied: sortedcontainers<3.0, >=2.0 in c:\users\firry\AppData\Local\Programs\Python\Python312\lib\site-packages (from intervaltree>=3->ipymarkup) (2.4.0)

In [60]:

```
from slovnet import NER
```

```
from ipymarkup import show_span_ascii_markup as show_markup
```

In [61]:

```
# Файл необходимо скачать по ссылке https://github.com/natasha/slovnet#downloads
```

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

```
ner_res = ner.navec(navec)
```

```
markup_ner = ner(text)
```

In [62]:

```
markup_ner
```

Out[62]:

```
SpanMarkup(
```

```
    text='Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе. Там разместят частную коллекцию произведений искусства. 4 этажа займут офисы, а верхний уровень отведен под ресторан с видовой террасой.'
```

```
    spans=[Span(
        start=67,
        stop=87,
        type='LOC'
    ),
```

```
    Span(
        start=90,
        stop=106,
        type='LOC'
    )
]
```

```
)
```

In [63]:

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

Пятнадцатизэтажный музейный комплекс с концертным залом построят на Малой Почтовой улице в Басманном районе. Там разместят частную

LOC_____ LOC_____

коллекцию произведений искусства. 4 этажа займут офисы, а верхний уровень отведен под ресторан с видовой террасой.

Разбор предложения

In [64]:

```
from natasha import NewsSyntaxParser
```

In [65]:

```
emb = NewsEmbedding()
```

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

In [66]:

```
n_doc.parse_syntax(syntax_parser)
```

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

