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

Лабораторная работа №5
По курсу «Методы машинного обучения»

«Предобработка текста»

ИСПОЛНИТЕЛЬ:

Лосева Светлана Сергеевна
Группа ИУ5-24М

_____

ПРОВЕРИЛ:

Гапанюк Ю.Е.

_____

Цель работы:
Изучение методов предобработки текста.

Задание:
Для произвольного предложения или текста решите следующие задачи:

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

Описание задания:
Для выполнения лабораторной работы возьмём фразу: «Предобработка данных в XML файле».

Выполнение работы:

1. Токенизация. NLTK
2. Частеречная разметка. Natasha
3. Лемматизация. Natasha
4. Выделение именованных сущностей. Natasha
5. Разбор предложения. Natasha

Вывод:
Была проделана работа по изучению методов предобработки текста, все задачи были выполнены.

```
!pip install numpy pandas scikit-surprise sklearn seaborn matplotlib spacy nltk navec slov
```

```
     |████████████████████████████████| 34.4MB 109kB/s
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: scipy>=1.0.0 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: six>=1.10.0 in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/lo
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in /usr/local/lib/python3.7/dis
Requirement already satisfied: thinc==7.4.0 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.7
Requirement already satisfied: plac<1.2.0,>=0.9.6 in /usr/local/lib/python3.7/dist
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.7/dis
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in /usr/local/lib/python3.7
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.7/d
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.7/dis
Requirement already satisfied: blis<0.5.0,>=0.4.0 in /usr/local/lib/python3.7/dist
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in /usr/local/lib/python3.7/di
Requirement already satisfied: razdel in /usr/local/lib/python3.7/dist-packages (f
Collecting ipymarkup>=0.8.0
  Downloading https://files.pythonhosted.org/packages/bf/9b/bf54c98d50735a4a7c84c7
Collecting yargy>=0.14.0
  Downloading https://files.pythonhosted.org/packages/d3/46/bc1a17200a55f4b0608f39
     |████████████████████████████████| 51kB 5.6MB/s
Collecting pymorphy2
  Downloading https://files.pythonhosted.org/packages/07/57/b2ff2fae3376d4f3c697b9
     |████████████████████████████████| 61kB 6.4MB/s
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packa
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/loc
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8" in
Collecting intervaltree>=3
  Downloading https://files.pythonhosted.org/packages/50/fb/396d568039d21344639db9
Collecting pymorphy2-dicts-ru<3.0,>=2.4
  Downloading https://files.pythonhosted.org/packages/3a/79/bea0021eeb7eeefde22ef9
     |████████████████████████████████| 8.2MB 19.5MB/s
Collecting dawg-python>=0.7.1
  Downloading https://files.pythonhosted.org/packages/6a/84/ff1ce2071d4c650ec85745
Requirement already satisfied: docopt>=0.6 in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages

Requirement already satisfied: typing-extensions>=3.6.4; python_version < "3.8" in
Requirement already satisfied: sortedcontainers<3.0,>=2.0 in /usr/local/lib/python
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
  Stored in directory: /root/.cache/pip/wheels/f3/f2/66/e9c30d3e9499e65ea2fa0d07c0
Successfully built intervaltree
Installing collected packages: intervaltree, ipymarkup, pymorphy2-dicts-ru, dawg-p
  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 natash
```

Токенизация. NLTK

```
import nltk
nltk.download('punkt')
text1 = 'Предобработка данных в XML файле.'
text2 = 'Меня зовут Бонд. Джеймс Бонд'
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
from nltk import tokenize
dir(tokenize)[:18]
```

```
['BlanklineTokenizer',
 'LineTokenizer',
 'MWETokenizer',
 'PunktSentenceTokenizer',
 'RegexpTokenizer',
 'ReppTokenizer',
 'SExprTokenizer',
 'SpaceTokenizer',
 'StanfordSegmenter',
 'TabTokenizer',
 'TextTilingTokenizer',
 'ToktokTokenizer',
 'TreebankWordTokenizer',
 'TweetTokenizer',
 'WhitespaceTokenizer',
 'WordPunctTokenizer',
 '__builtins__',
 '__cached__']
```

```
nltk_tk_1 = nltk.WordPunctTokenizer()
nltk_word = nltk_tk_1.tokenize(text1)
print(nltk_word)
```

```
['Предобработка', 'данных', 'в', 'XML', 'файле', '.']
```

```
# Токенизация по предложениям
nltk_tk_sents = nltk.tokenize.sent_tokenize(text1)
print(len(nltk_tk_sents))
nltk_tk_sents
```

```
1
['Предобработка данных в XML файле.']
```

Частеречная разметка. Natasha

```
from navec import Navec
from slovnet import Morph


from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
navec = Navec.load('/content/gdrive/My Drive/MMO/navec_news_v1_1B_250K_300d_100q.tar')
n_morph = Morph.load('/content/gdrive/My Drive/MMO/slovnet_morph_news_v1.tar', batch_size=


morph_res = n_morph.navec(navec)


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


n_text1_markup = list(_ for _ in n_morph.map(nltk_tk_sents))
[print_pos(x) for x in n_text1_markup]
```

```
П - PROPN|Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
р - NOUN
е - X|Foreign=Yes
д - NOUN
о - X|Foreign=Yes
б - NOUN|Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
р - X|Foreign=Yes
а - CCONJ
б - PROPN
о - NOUN|Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
т - PRON|Animacy=Inan|Case=Loc|Gender=Neut|Number=Sing
к - ADP
а - X|Foreign=Yes
  - PUNCT
д - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
а - CCONJ
н - X|Foreign=Yes
н - X|Foreign=Yes
ы - X|Foreign=Yes
х - X|Foreign=Yes
  - PUNCT
в - X|Foreign=Yes
  - PUNCT
Х - X|Foreign=Yes
М - PROPN|Foreign=Yes
L - X|Foreign=Yes
  - PUNCT
ф - X|Foreign=Yes
а - CCONJ
й - ADJ|Case=Nom|Degree=Pos|Gender=Masc|Number=Sing
л - X|Foreign=Yes
е - NOUN|Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
. - PUNCT
[None]
```

Лемматизация. Natasha

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


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_doc1 = n_lemmatize(text1)
{_.text: _.lemma for _ in n_doc1.tokens}
```

```
{'.': '.',
 'XML': 'xml',
 'Предобработка': 'предобработка',
 'в': 'в',
 'данных': 'данные',
 'файле': 'файл'}
```

```
n_doc2 = n_lemmatize(text2)
{_.text: _.lemma for _ in n_doc2.tokens}
```

```
{'.': '.', 'Бонд': 'бонд', 'Джеймс': 'джеймс', 'Меня': 'я', 'зовут': 'звать'}
```

Выделение (распознавание) именованных сущностей. Natasha

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


ner = NER.load('/content/gdrive/My Drive/MMO/slovnet_ner_news_v1.tar')


ner_res = ner.navec(navec)


markup_ner2 = ner(text2)


markup ner2
```

```
'—

SpanMarkup(
    text='Меня зовут Бонд. Джеймс Бонд',
    spans=[Span(
         start=11,
         stop=15,
         type='PER'
     ), Span(
         start=17,
         stop=28,
         type='PER'
     )]
)
```

```
show_markup(markup_ner2.text, markup_ner2.spans)
```

```
Меня зовут Бонд. Джеймс Бонд
           PER—  PER————————
```

Разбор предложения. Natasha

```
from natasha import NewsSyntaxParser
```

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

```
n_doc1.parse_syntax(syntax_parser)
n_doc1.sents[0].syntax.print()
```

```
┌► Предобработка amod
└─ данных
┌► в             case
└─ XML
   файле
   .
```

```
n_doc2.parse_syntax(syntax_parser)
n_doc2.sents[0].syntax.print()
```

```
      ┌► Меня  obj
┌┌─┌─└─ зовут
│└─└──► Бонд  xcomp
└────►  .     punct
```