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

Лабораторная работа №8
по дисциплине
«Методы машинного обучения»
на тему

«Предобработка текста»

Выполнил:
студент группы ИУ5И-22М
Лу Жуньда

Москва — 2024 г.

1. Цель лабораторной работы

Изучение методов предобработки текстов.

2. Задание

Для произвольного предложения или текста решите следующие задачи:

1. Токенизация.
2. Частеречная разметка.
3. Лемматизация.
4. Выделение (распознавание) именованных сущностей.
5. Разбор предложения.

3. Текст программы

Для выполнения задач предобработки текста на примере произвольного предложения, мы воспользуемся библиотекой nltk и spacy в Python.

```
!pip install nltk spacy
!python -m spacy download ru_core_news_md
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: spacy in /usr/local/lib/python3.10/dist-packages (3.7.4)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2024.5.15)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.4)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.3.0,>=8.2.2 in /usr/local/lib/python3.10/dist-packages (from spacy) (8.2.3)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.10/dist-packages (from spacy) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (0.3.4)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (0.9.4)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy) (6.4.0)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (2.31.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (from spacy) (2.7.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy) (3.1.4)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy) (67.7.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (24.0)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (3.4.0)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.10/dist-packages (from spacy) (1.25.2)
```

1. Токенизация

Токенизация заключается в разбиении текста на отдельные слова (токены).

```
[2] import nltk
    nltk.download('punkt')

    text = "Иван Иванович Петров отправился в Москву на конференцию по обработке естественного языка."
    tokens = nltk.word_tokenize(text, language='russian')
    print("Токены:", tokens)
```

```
Токены: ['Иван', 'Иванович', 'Петров', 'отправился', 'в', 'Москву', 'на', 'конференцию', 'по', 'обработке', 'естественного', 'языка', '.']
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

2. Частеречная разметка

Частеречная разметка (POS-tagging) заключается в определении частей речи для каждого слова.

```
nltk.download('averaged_perceptron_tagger_ru')

pos_tags = nltk.pos_tag(tokens, lang='rus')
print("Части речи:", pos_tags)
```

```
[nltk_data] Downloading package averaged_perceptron_tagger_ru to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger_ru is already up-to-
[nltk_data]       date!
Части речи: [('Иван', 'S'), ('Иванович', 'S'), ('Петров', 'S'), ('отправился', 'V'), ('в', 'PR'), ('Москву', 'S'), ('на', 'PR'), ('конференцию', 'S'), ('по', 'PR'), ('обработке', 'S'), ('естественного', 'A=m'), ('языка', 'S'), ('.', 'NONLEX')]
```

3. Лемматизация

Лемматизация заключается в приведении слов к их начальной форме (лемме).

```
import spacy

nlp = spacy.load("ru_core_news_md")
doc = nlp(text)

lemmas = [token.lemma_ for token in doc]
print("Леммы:", lemmas)
```

```
Леммы: ['иван', 'иванович', 'петров', 'отправиться', 'в', 'москва', 'на', 'конференция', 'по', 'обработка', 'естественный', 'язык', '.']
```

4. Выделение (распознавание) именованных сущностей

Распознавание именованных сущностей (NER) заключается в идентификации имен людей, организаций, мест и т.д.

```
[5] entities = [(entity.text, entity.label_) for entity in doc.ents]
    print("Именованные сущности:", entities)
```

```
Именованные сущности: [('Иван Иванович Петров', 'PER'), ('Москву', 'LOC')]
```

5. Разбор предложения

Разбор предложения заключается в анализе синтаксической структуры предложения.

```
[6] for token in doc:
        print(f'{token.text} ({token.dep_}) --> {token.head.text}')
```

```
Иван (nsubj) --> отправился
Иванович (appos) --> Иван
Петров (appos) --> Иван
отправился (ROOT) --> отправился
в (case) --> Москву
Москву (obl) --> отправился
на (case) --> конференцию
конференцию (obl) --> отправился
```