МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа № 8
по дисциплине «Методы машинного обучения»

Тема: « Предобработка текста»

ИСПОЛНИТЕЛЬ: Ли Яцзинь
ФИО
группа ИУ5И-21М
подпись"

ПРЕПОДАВАТЕЛЬ: Гапанюк Ю .Е

Москва - 2024

Задание:

Для произвольного предложения или текста решите следующие задачи:

Токенизация.

Частеречная разметка.

Лемматизация.

Выделение (распознавание) именованных сущностей.

Разбор предложения.

текст программы

```
!pip install spacy
!python -m spacy download en_core_web_sm

import spacy

# Загрузка модели SpaCy для английского языка
nlp = spacy.load("en_core_web_sm")

# Произвольное предложение для анализа
sentence = "The quick brown fox jumps over the lazy dog."

# Токенизация
doc = nlp(sentence)
tokens = [token.text for token in doc]
print("Токены:", tokens)
```

```
# Частеречная разметка
pos_tags = [(token.text, token.pos_) for token in doc]
print("Частеречная разметка:", pos_tags)

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

# Выделение именованных сущностей
entities = [(entity.text, entity.label_) for entity in doc.ents]
print("Именованные сущности:", entities)

# Разбор предложения
parsing = [(token.text, token.dep_) for token in doc]
print("Разбор предложения:", parsing
```

экранные формы с примерами выполнения программы.

```
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2024.6.2)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.1.5)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.10.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (8.1.7)
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in /usr/local/lib/python3.10/dist-packages (from weasel<0.4.0,>=0.1.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (0.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (2.1.5)
Requirement already satisfied: marisa-trie>=0.7.7 in /usr/local/lib/python3.10/dist-packages (from language-data>=1.2->langcodes<4.0.0,>=3.2.0->spacy<3.8.0,>=3.7.2->en-core-web-sm==3.7.1) (1.1.1)
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
⚠ Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
Токены: ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog', '.']
Частеречная разметка: [('The', 'DET'), ('quick', 'ADJ'), ('brown', 'ADJ'), ('fox', 'NOUN'), ('jumps', 'VERB'), ('over', 'ADP'), ('the', 'DET'), ('lazy', 'ADJ'), ('dog', 'NOUN'), ('.', 'PUNCT')]
Леммы: ['the', 'quick', 'brown', 'fox', 'jump', 'over', 'the', 'lazy', 'dog', '.']
Именованные сущности: []
Разбор предложения: [('The', 'det'), ('quick', 'amod'), ('brown', 'amod'), ('fox', 'nsubj'), ('jumps', 'ROOT'), ('over', 'prep'), ('the', 'det'), ('lazy', 'amod'), ('dog', 'pobj'), ('.', 'punct')]
```

Рисунок 3. результаты