Министерство науки и высшего образования
Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
"Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)"
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА ___СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)_____

Отчет по лабораторной работе №4

«Предобработка текста»

по курсу «Методы машинного обучения».

ИСПОЛНИТЕЛЬ: Группа ИУ5-24М
Уралова Е.А.
ФИО
_____
подпись

"26" апреля 2024 г.

ПРЕПОДАВАТЕЛЬ: Гапанюк Ю.Е.
ФИО
_____
подпись

"__"_____2024 г.

Москва – 2024

Задание:

Для произвольного предложения или текста решите следующие задачи:

1) Токенизация.

2) Частеречная разметка.

3) Лемматизация.

4) Выделение (распознавание) именованных сущностей.

5) Разбор предложения.

Выполнение:

+ Код + Текст

```
[6]  import spacy
```

Загрузка модели языка:

```
[7]  nlp = spacy.load("en_core_web_sm")
```

Пример текста:

```
[14] text = "Кошка спит на подоконнике, пока ребенок делает задание."
```

Токенизация:

```
[15] doc = nlp(text)
     tokens = [token.text for token in doc]
     print("Токенизация:")
     print(tokens)
     print()
```

```
Токенизация:
['Кошка', 'спит', 'на', 'подоконнике', ',', 'пока', 'ребенок', 'делает', 'задание', '.']
```

Частеречная разметка:

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

```
Частеречная разметка:
[('Кошка', 'PROPN'), ('спит', 'NOUN'), ('на', 'PROPN'), ('подоконнике', 'PROPN'), (',', 'PUNCT'), ('пока', 'NOUN'), ('ребенок', 'ADJ'), ('делает
```

Лемматизация:

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

```
Лемматизация:
['Кошка', 'спит', 'на', 'подоконнике', ',', 'пока', 'ребенок', 'делает', 'задание', '.']
```

Выделение (распознавание) именованных сущностей:

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

```
Именованные сущности:
[('Кошка', 'PERSON'), ('задание', 'PERSON')]
```

Разбор предложения:

```
[19] parsed_sentences = [chunk.text for chunk in doc.noun_chunks]
     print("Разбор предложения:")
     print(parsed_sentences)
```

```
Разбор предложения:
['Кошка спит на подоконнике', 'пока ребенок делает задание']
```

Вывод:

В данной лабораторной работе рассмотрено решение следующих задач: токенизация, частеречная разметка, лемматизация, выделение (распознавание) именованных сущностей, разбор предложения.