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 и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

Лабораторная работа №6 по курсу «Методы машинного обучения в автоматизированных системах обработки информации и управления»

Подготовили:

У Жун

ИУ5И-25М

01.06.2024

Проверил:

Гапанюк Ю. Е.

2024 г.

Цель лабораторной работы:
изучение методов предобработки текстов.

Задание:
Для произвольного предложения или текста решите следующие задачи:

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

Препроцессинг русского текста с помощью spaCy для предложения "Лето в южном Китае жаркое":

Установка и загрузка модели spaCy

```
import spacy
nlp = spacy.load('ru_core_news_sm')
```

Рис.1-Установка и загрузка модели spaCy

Токенизация

```
text = "Летом на юге Китая жарко."
doc = nlp(text)

tokens = [token.text for token in doc]
print("Tokens:", tokens)
```

```
Tokens: ['Летом', 'на', 'юге', 'Китая', 'жарко', '.']
```

Рис.2-Токенизация

Частеречная разметка

```
pos_tags = [(token.text, token.pos_) for token in doc]
print("POS tags:", pos_tags)
```

```
POS tags: [('Летом', 'NOUN'), ('на', 'ADP'), ('юге', 'NOUN'),
('Китая', 'PROPN'), ('жарко', 'ADJ'), ('.', 'PUNCT')]
```

Рис.3-Частеречная разметка

Лемматизация

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

```
Lemmas: [('Летом', 'лето'), ('на', 'на'), ('юге', 'юг'), ('Китая', 'китай'), ('жарко', 'жаркий'), ('.', '.')]
```

Рис.4-Лемматизация

Выделение именованных сущностей

```
entities = [(ent.text, ent.label_) for ent in doc.ents]
print("Entities:", entities)
```

```
Entities: [('Китая', 'LOC')]
```

Рис.5-Выделение именованных сущностей

Разбор предложения

```
print("Dependency parse:")
for token in doc:
    print(f'{token.text} -> {token.dep_} -> {token.head.text}')
```

```
Dependency parse:
Летом -> obl -> жарко
на -> case -> юге
юге -> nmod -> Летом
Китая -> nmod -> юге
жарко -> ROOT -> жарко
. -> punct -> жарко
```

Рис.6-Разбор предложения