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

Лабораторная работа № 7

По курсу «методы машинного обучения в АСОИУ»

«Предобработка текста»

Выполнил:

студент ИУ5-23М
Семенов И.А.

Проверил:

Балашов А.М.

Подпись:

29.05.2024

Москва, 2024

Описание задания

Для произвольного предложения или текста решите следующие задачи:

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

Описание задач

Частеречная разметка текста (также известная как POS-tagging или Part-of-Speech tagging) — это процесс определения грамматических категорий (частей речи) слов в тексте. В качестве таких категорий могут выступать существительные, глаголы, прилагательные и так далее.

Выделение (распознавание) именованных сущностей (NER, Named Entity Recognition) — это процесс распознавания и классификации слов или фраз в тексте в определенные категории, такие как имена людей, названия мест, организации и так далее.

Текст программы и экранные формы с примерами выполнения программы

Токенизация

```
# Импортируем необходимые библиотеки
import nltk
nltk.download('punkt')  # Загрузим ресурс для токенизации

# Введем текст для токенизации
text = "Это пример предложения для демонстрации токенизации. Это очень полезная задача для обработки естественного языка."

# Токенизация текста
tokens = nltk.word_tokenize(text)

# Вывод токенов
print("Токены:", tokens)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Токены: ['Это', 'пример', 'предложения', 'для', 'демонстрации', 'токенизации', '.', 'Это', 'очень', 'полезная', 'задача', 'для', 'обработки', 'естественного', 'языка', '.']
```

Частеречная разметка

```python
# Импортируем spaCy и загружаем модель для русского языка
import spacy

# Загрузим модель для русского языка
nlp = spacy.load("ru_core_news_sm")

# Пример текста для разметки
text = "Это пример предложения для демонстрации частеречной разметки. Это очень полезная задача для обработки естественного языка."

# Анализ текста
doc = nlp(text)

# Проведение частеречной разметки
for token in doc:
    print(f"Слово: {token.text}, Часть речи: {token.pos_}")
```

```
Слово: Это, Часть речи: PRON
Слово: пример, Часть речи: NOUN
Слово: предложения, Часть речи: NOUN
Слово: для, Часть речи: ADP
Слово: демонстрации, Часть речи: NOUN
Слово: частеречной, Часть речи: ADJ
Слово: разметки, Часть речи: NOUN
Слово: ., Часть речи: PUNCT
Слово: Это, Часть речи: PRON
Слово: очень, Часть речи: ADV
Слово: полезная, Часть речи: ADJ
Слово: задача, Часть речи: NOUN
Слово: для, Часть речи: ADP
Слово: обработки, Часть речи: NOUN
Слово: естественного, Часть речи: ADJ
Слово: языка, Часть речи: NOUN
Слово: ., Часть речи: PUNCT
```

Лемматизация

```python
# Импортируем spaCy и загружаем модель для русского языка
import spacy

# Загрузим модель для русского языка
nlp = spacy.load("ru_core_news_sm")

# Пример текста для лемматизации
text = "Это пример предложения для демонстрации лемматизации. Это очень полезная задача для обработки естественного языка."

# Анализ текста
doc = nlp(text)

# Выполним лемматизацию и выведем результат
for token in doc:
    print(f"Слово: {token.text}, Лемма: {token.lemma_}")
```

```
Слово: Это, Лемма: это
Слово: пример, Лемма: пример
Слово: предложения, Лемма: предложение
Слово: для, Лемма: для
Слово: демонстрации, Лемма: демонстрация
Слово: лемматизации, Лемма: лемматизация
Слово: ., Лемма: .
Слово: Это, Лемма: это
Слово: очень, Лемма: очень
Слово: полезная, Лемма: полезный
Слово: задача, Лемма: задача
Слово: для, Лемма: для
Слово: обработки, Лемма: обработка
Слово: естественного, Лемма: естественный
Слово: языка, Лемма: язык
Слово: ., Лемма: .
```

• Выделение (распознавание) именованных сущностей.

```
[14] # Импортируем spaCy и загружаем модель для русского языка
     import spacy

     # Загрузим модель для русского языка
     nlp = spacy.load("ru_core_news_sm")

     # Пример текста для распознавания именованных сущностей
     text = "Иван Иванов работает в Google в Москве."

     # Анализ текста
     doc = nlp(text)

     # Выделение именованных сущностей
     print("Именованные сущности в тексте:")
     for ent in doc.ents:
         print(f"Текст: {ent.text}, Тип: {ent.label_}")
```

```
Именованные сущности в тексте:
Текст: Иван Иванов, Тип: PER
Текст: Google, Тип: ORG
Текст: Москве, Тип: LOC
```

Разбор предложения.

```
# Импортируем spaCy и загружаем модель для русского языка
import spacy

# Загрузим модель для русского языка
nlp = spacy.load("ru_core_news_sm")

# Пример текста для синтаксического анализа
text = "Иван Иванов работает в Google в Москве."

# Анализ текста
doc = nlp(text)

# Выведем слова и их синтаксические зависимости
print("Синтаксический анализ предложения:")
for token in doc:
    print(f"Слово: {token.text}, Синтаксическая зависимость: {token.dep_}, Родитель: {token.head.text}")
```

```
Синтаксический анализ предложения:
Слово: Иван, Синтаксическая зависимость: nsubj, Родитель: работает
Слово: Иванов, Синтаксическая зависимость: appos, Родитель: Иван
Слово: работает, Синтаксическая зависимость: ROOT, Родитель: работает
Слово: в, Синтаксическая зависимость: case, Родитель: Google
Слово: Google, Синтаксическая зависимость: obl, Родитель: работает
Слово: в, Синтаксическая зависимость: case, Родитель: Москве
Слово: Москве, Синтаксическая зависимость: obl, Родитель: работает
Слово: ., Синтаксическая зависимость: punct, Родитель: работает
```

Вывод

В рамках данной лабораторной работы была проведена токенезация, частеречная разметка, лемматизация, выделение (распознавание) именованных сущностей.