

9. Практическое занятие: Извлечение сущностей (NER) и анализ тональности (Sentiment Analysis)

Цель занятия

Научиться извлекать из текста структурированную информацию (имена, организации, локации) и определять эмоциональную окраску высказываний с использованием готовых предобученных моделей.

Теоретический минимум

- 1. NER (Named Entity Recognition):** задача идентификации и классификации сущностей в тексте. Позволяет ответить на вопросы: «Кто?», «Где?», «Когда?», «Какая компания?».
- 2. Sentiment Analysis:** задача классификации текстов по эмоциональному признаку (позитивный, негативный, нейтральный). Помогает автоматизировать обработку отзывов и мониторинг соцсетей.

Задание 1. Распознавание сущностей (NER) с помощью Natasha

Для русского языка библиотека Natasha является одним из самых эффективных и быстрых инструментов NER.

Инструкция:

Обработайте новостной текст и извлеките из него имена людей и названия организаций.

```
from natasha import (
    Segmenter, MorphVocab, NewsEmbedding,
    NewsNERTagger, Doc
)

# Инициализация инструментов
segmenter = Segmenter()
morph_vocab = MorphVocab()
emb = NewsEmbedding()
ner_tagger = NewsNERTagger(emb)

text = "Антон Силуанов заявил, что Министерство финансов и Сбербанк согласовали условия кредитования в Москве."
doc = Doc(text)

# Обработка
doc.segment(segmenter)
doc.tag_ner(ner_tagger)

print(f'{ "Сущность":<25} | { "Тип"}')
for span in doc.spans:
    print(f'{span.text:<25} | {span.type}')
```

Задание 2. Анализ тональности (Sentiment Analysis)

Для оценки эмоционального фона мы воспользуемся моделью на базе библиотеки `dostoevsky` или предобученными трансформерами из Hugging Face.

Инструкция:

Проанализируйте список отзывов и определите их полярность.

```
# Пример с использованием упрощенной логики или библиотеки dostoevsky
# pip install dostoevsky
from dostoevsky.tokenization import RegexTokenizer
from dostoevsky.models import FastTextSocialNetworkModel

tokenizer = RegexTokenizer()
model = FastTextSocialNetworkModel(tokenizer=tokenizer)

messages = [
    'Этот курс по NLP просто потрясающий!',
    'Ужасный сервис, я очень разочарован.',
    'Обычный телефон, ничего особенного.'
]

results = model.predict(messages, k=2)

for message, sentiment in zip(messages, results):
    # Берем наиболее вероятную категорию
    label = max(sentiment, key=sentiment.get)
    print(f"Текст: {message}")
    print(f"Тональность: {label} (уверенность: {sentiment[label]:.2f})\n")
```

Задание 3. Визуализация результатов NER

Визуальное представление сущностей помогает быстрее проверить точность работы модели.

Инструкция:

Используйте встроенные средства визуализации (если доступно в среде) или выведите текст, подсветив сущности типа LOC (локации).

```
# Логика: если тип сущности LOC, вывести её заглавными буквами
modified_text = text
for span in reversed(doc.spans):
    if span.type == 'LOC':
        modified_text = modified_text[:span.start] +
f" [{span.text.upper()}]" + modified_text[span.end:]

print("Текст с выделенными локациями:", modified_text)
```

Задание 4. Анализ связи сущностей и тональности

Часто бизнес-задача звучит так: «Что клиенты говорят именно о нашем бренде?». Это требует объединения NER и Sentiment Analysis.

Инструкция:

Дан текст: "Вчера купил кроссовки Nike в магазине Спортмастер. Nike очень удобные, а вот Спортмастер расстроил медленной доставкой".

1. Найдите все упоминания бренда (NER).
2. Определите тональность предложений, где встречаются эти бренды.
3. Сделайте вывод: к какой сущности относится негатив, а к какой — позитив.

Контрольные вопросы

1. В чем разница между тегами LOC (местоположение) и ORG (организация)?
2. Почему анализ тональности может ошибаться на саркастических фразах (например, "*O да, лучший сервис в мире (нет)*")?
3. Как NER помогает в автоматизации заполнения юридических документов?

Итог работы

Студенты научились переходить от статистических свойств текста к извлечению фактов и эмоций. Эти навыки являются ключевыми для создания чат-ботов, систем мониторинга репутации и интеллектуальных поисковых систем.