

## 6. Практическое занятие: Частеречная разметка (POS-tagging) и работа со стоп-словами

### Цель занятия

Научиться определять части речи слов в предложении (Part-of-Speech tagging) для извлечения смысловых конструкций и освоить методы фильтрации текста от стоп-слов для повышения качества анализа.

### Теоретический минимум

1. **POS-tagging:** процесс присвоения слову морфологического признака (существительное, глагол, прилагательное и т.д.) на основе его определения и контекста. Это критично для разрешения неоднозначности (например, слово «стекло» может быть существительным или глаголом).
2. **Стоп-слова:** часто встречающиеся слова (предлоги, союзы, местоимения), которые обычно не несут самостоятельного смысла и удаляются для уменьшения объема данных и выделения ключевых слов.

### Задание 1. Определение частей речи с помощью PyMorphy2

В русском языке для автоматической разметки часто используется библиотека `rumorphy2`. Она предоставляет подробную информацию о граммемах слова.

#### Инструкция:

Проанализируйте предложение и выведите части речи для каждого слова.

#### Python

```
import rumorphy2

morph = rumorphy2.MorphAnalyzer()
sentence = "Интеллектуальные системы быстро обрабатывают сложные тексты"
tokens = sentence.lower().split()

print("Результат POS-разметки:")
for word in tokens:
    p = morph.parse(word)[0]
    # tag.POS возвращает аббревиатуру части речи (NOUN, VERB, ADJF и т.д.)
    print(f"{word} -> {p.tag.POS} ({p.tag.cyr_repr})")
```

### Задание 2. Разрешение морфологической неоднозначности

Слово «стали» может быть глаголом (они стали) или существительным (прочность стали).

#### Инструкция:

Проверьте, как `rumorphy2` предлагает варианты разбора для слова «стали».

## Python

```
word_ambiguous = "стали"
variants = morph.parse(word_ambiguous)

for i, v in enumerate(variants):
    print(f"Вариант {i+1}: Форма: {v.normal_form}, Тер: {v.tag.POS}")
```

*Примечание: Поскольку PyMorphy2 работает с отдельными словами, он выдает все возможные варианты. Для учета контекста в NLP используют более сложные модели (например, SpaCy или Stanza).*

## Задание 3. Фильтрация стоп-слов

Для удаления «информационного шума» используются заранее подготовленные списки стоп-слов.

### Инструкция:

Загрузите стандартный список стоп-слов из библиотеки nltk и очистите от них текст.

## Python

```
import nltk
from nltk.corpus import stopwords

# Загрузка набора стоп-слов
nltk.download('stopwords')
russian_stopwords = stopwords.words('russian')

text = "Вчера я пошел в магазин и купил там очень вкусное яблоко."
words = text.lower().split()

# Очистка текста
filtered_words = [word for word in words if word not in russian_stopwords]

print("Список стоп-слов (первые 10):", russian_stopwords[:10])
print("Текст после очистки:", " ".join(filtered_words))
```

## Задание 4. Извлечение только значимых частей речи

Иногда для анализа тематики текста (например, построения облака тегов) нужно оставить только существительные и прилагательные.

### Инструкция:

Напишите функцию, которая фильтрует текст, оставляя только слова определенных категорий (например, NOUN и ADJF).

## Python

```
def extract_key_content(text):
```

```

tokens = text.lower().split()
target_pos = {'NOUN', 'ADJF'}
result = []

for word in tokens:
    p = morph.parse(word)[0]
    if p.tag.POS in target_pos:
        result.append(p.normal_form)
return result

sample_text = "Быстрый искусственный интеллект создает удивительные  
возможности для развития науки"
print("Только важные леммы:", extract_key_content(sample_text))

```

## Задание 5. Анализ специфических стоп-слов

Списки стоп-слов зависят от задачи. Например, в юридических текстах слово «закон» встречается часто, но не является стоп-словом.

Инструкция:

Добавьте в список стоп-слов свои собственные (например, название вашей компании или специфические вводные слова) и проведите повторную очистку.

Python

```

custom_stops = ["интеллект", "возможности"]
extended_stopwords = russian_stopwords + custom_stops
# Примените этот список к тексту из Задания 4

```

## Контрольные вопросы

1. Какую информацию, кроме части речи, можно извлечь из тега в PyMorphy2? (Падеж, число, род).
2. Почему в задачах классификации отзывов (позитивные/негативные) опасно удалять частицу «не»?
3. Что такое «ложные стоп-слова» и как они могут повлиять на смысл предложения?

Итог работы

Обучающиеся получили практический навык морфологической разметки и очистки текста. Эти методы позволяют сфокусироваться на семантическом ядре текста, отсекая служебные части речи и грамматический «шум».