12. Практическое занятие: Стиллометрия. Определение стилистики статистическими и лингвистическими методами

Цель занятия

Изучить методы количественного анализа авторского стиля (стиллометрии), научиться выявлять «цифровой отпечаток» автора через статистические характеристики текста и лингвистические особенности употребления слов.

Теоретический минимум

Стиллометрия — это область лингвистики, основанная на предположении, что у каждого автора есть уникальные, часто неосознаваемые привычки письма. К ним относятся:

1. **Служебные слова:** частота употребления предлогов, союзов и частиц (автор почти не может их контролировать).
2. **Богатство словаря (TTR):** отношение количества уникальных слов к общему числу слов в тексте.
3. **Длина предложений и слов:** распределение длин структурных единиц.
4. **Закон Ципфа:** статистическое распределение частотности слов в языке.

Задание 1. Расчет коэффициента лексического разнообразия (TTR)

Коэффициент Type-Token Ratio (TTR) показывает, насколько богат словарь автора.

$$TTR = \frac{\text{Количество уникальных слов (Types)}}{\text{Общее количество слов (Tokens)}}$$

Инструкция:

Сравните два текста и определите, у какого автора лексикон более разнообразен.

Python

```python
def calculate_ttr(text):
    tokens = re.sub(r'[^а-яё\s]', '', text.lower()).split()
    if not tokens: return 0
    types = set(tokens)
    return len(types) / len(tokens)

author1 = "Ночь, улица, фонарь, аптека, Бессмысленный и тусклый свет."
author2 = "Вчера я пошел в магазин и купил там вчерашний хлеб и вчерашнее молоко."

print(f"TTR Автор 1 (Блок): {calculate_ttr(author1):.2f}")
print(f"TTR Автор 2: {calculate_ttr(author2):.2f}")
```

Задание 2. Анализ распределения длин предложений

Стиль часто характеризуется ритмикой. Длинные, сложные предложения характерны для классической прозы, короткие — для новостей или социальных сетей.

Инструкция:

Напишите функцию, которая вычисляет среднюю длину предложения в словах.

Python
```python
import numpy as np

def sentence_length_stats(text):
    sentences = re.split(r'[.!?]+', text)
    sentences = [s.strip() for s in sentences if len(s.strip()) > 0]
    lengths = [len(s.split()) for s in sentences]
    return {
        "mean": np.mean(lengths),
        "max": np.max(lengths),
        "std": np.std(lengths) # Вариативность стиля
    }

text_sample = "Это очень длинное предложение, которое содержит много слов и знаков. Это коротко. Мы пишем код."
print("Статистика предложений:", sentence_length_stats(text_sample))
```

Задание 3. Использование служебных слов (Функциональные слова)

Стиль автора лучше всего определяют не «умные» слова, а то, как он использует «и», «но», «в», «на».

Инструкция:

Выделите топ-10 самых частых слов в двух разных текстах, включая стоп-слова. Сравните их частотные профили.

Python
```python
from collections import Counter

def get_style_profile(text, top_n=10):
    tokens = re.sub(r'[^а-яё\s]', '', text.lower()).split()
    return Counter(tokens).most_common(top_n)

# Задание: сравните профили научного текста и личного письма
```

Задание 4. Метод «Дельта» Берроуза (Burrows' Delta)

Это один из самых известных методов в стиллометрии для определения авторства. Он измеряет разницу между частотами самых употребительных слов в неизвестном тексте и текстах известных авторов.

Инструкция (теоретическая):

Представьте, что автор А использует союз «и» с частотой 5%, а автор Б — 3%. Если в анонимном тексте частота «и» составляет 4.8%, к какому автору ближе этот текст по данному признаку?

Задание 5. Визуализация «стилистического отпечатка»

Постройте график распределения длин слов для двух текстов.

Python

```python
import matplotlib.pyplot as plt

def plot_word_length_dist(text, label):
    tokens = re.sub(r'[^а-яё\s]', '', text.lower()).split()
    lengths = [len(w) for w in tokens]
    plt.hist(lengths, bins=range(1, 20), alpha=0.5, label=label,
density=True)

# plot_word_length_dist(text1, "Пушкин")
# plot_word_length_dist(text2, "Толстой")
# plt.legend()
# plt.show()
```

Контрольные вопросы

1. Почему при определении авторства стоп-слова важнее, чем существительные?
2. Как на коэффициент TTR влияет объем текста? (Почему TTR падает при увеличении длины текста?).
3. Может ли стиллометрия помочь в выявлении плагиата или генерации текста нейросетью?

Итог работы

Вы научились смотреть на текст как на набор статистических закономерностей. Стиллометрия позволяет не только атрибутировать анонимные тексты, но и анализировать эволюцию стиля писателя, а также отличать тексты, написанные человеком, от сгенерированных ИИ.