11. Практическое занятие: Тематическое моделирование (Topic Modelling) и прогнозирование

Цель занятия

Освоить методы автоматического выделения тем в коллекции текстов (LDA), провести глубокий лингвистический анализ полученных результатов и построить базовую модель для прогнозирования категорий текста.

Теоретический минимум

1. **LDA (Latent Dirichlet Allocation):** статистический метод, который представляет каждый документ как смесь нескольких тем, а каждую тему — как распределение слов с определенной вероятностью.
2. **Лингвистический анализ:** интерпретация ключевых слов темы для определения ее смысла (например, слова «банк», «процент», «кредит» формируют тему «Финансы»).
3. **Прогностическая модель:** использование вероятностей тем как признаков (features) для предсказания свойств документа (например, предсказание рейтинга отзыва на основе его тематики).

Задание 1. Тематическое моделирование с помощью LDA

Мы будем использовать библиотеку gensim для поиска скрытых тем в наборе документов.

Инструкция:

Подготовьте корпус, создайте словарь и обучите модель LDA на 2 темы.

```
from gensim import corpora
from gensim.models import LdaModel
import re

# Пример данных
documents = [
    "Технологии искусственного интеллекта развиваются быстро",
    "Процессоры и видеокарты стали мощнее в этом году",
    "Рецепт домашнего хлеба в духовке очень прост",
    "Кулинария требует внимания к деталям и специям"
]

# Предобработка: токенизация и очистка
texts = [re.sub(r'[^а-яё\s]', '', doc.lower()).split() for doc in
documents]

# Создание словаря и корпуса (мешок слов)
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]
```

```
# Обучение LDA
lda_model = LdaModel(corpus=corpus, id2word=dictionary, num_topics=2,
passes=10)

# Вывод тем
for idx, topic in lda_model.print_topics(-1):
    print(f"Тема №{idx}: {topic}")
```

Задание 2. Визуализация и лингвистическая интерпретация

Важно не просто получить список слов, но и понять, какую концепцию они описывают.

Инструкция:

Проанализируйте вывод Задания 1.

1. Выпишите 3 ключевых слова для каждой темы.
2. Дайте название каждой теме (например, «Hardware/IT» и «Cooking»).
3. Найдите документ, который имеет смешанную тематику (если бы в тексте говорилось о «кулинарном приложении для смартфона»).

Задание 3. Извлечение тематических признаков для прогнозирования

Теперь мы превратим текст в вектор вероятностей тем. Это сжатое представление текста, которое можно подать на вход прогностической модели.

Инструкция:

Преобразуйте новый документ в вектор вероятностей тем.

```
new_doc = "Новые кулинарные технологии на базе ИИ"
new_bow = dictionary.doc2bow(re.sub(r'[^а-яё\s]', '',
new_doc.lower()).split())

topic_distribution = lda_model.get_document_topics(new_bow)
print(f"Распределение тем для фразы: {topic_distribution}")
```

Задание 4. Построение прогностической модели

Представим, что наша задача — спрогнозировать, будет ли статья популярной, основываясь на её темах.

Инструкция:

Используйте вероятности тем как входные данные (X) для простой логистической регрессии.

```
import numpy as np
from sklearn.linear_model import LogisticRegression

# Имитация признаков (вероятности Темы 0) и целевой переменной (1 - популярно, 0 - нет)
X = np.array([[0.9], [0.8], [0.1], [0.2]]) # Вероятности темы "Технологии"
y = np.array([1, 1, 0, 0]) # Техно-статьи сейчас популярны

clf = LogisticRegression().fit(X, y)

# Прогноз для нового документа из Задания 3
prob_topic_0 = dict(topic_distribution).get(0, 0)
prediction = clf.predict([[prob_topic_0]])
print(f"Прогноз популярности (1-да, 0-нет): {prediction[0]}")
```

Задание 5. Глубокий лингвистический анализ ошибок

Модели тематического моделирования часто ошибаются из-за омонимии или специфического жаргона.

Инструкция:

Объясните, как слово «язык» (программирования vs анатомический/кулинарный) может исказить результаты LDA. Предложите лингвистический способ решения этой проблемы (например, использование биграмм или выделение именованных сущностей перед обучением LDA).

Контрольные вопросы

1. Чем тематическое моделирование (LDA) отличается от классификации текстов? (Подсказка: обучение с учителем vs без учителя).
2. Зачем нужно удалять стоп-слова перед запуском LDA?
3. Что такое когерентность (coherence score) темы и почему она важна для лингвиста?

Итог работы

Вы научились не просто классифицировать тексты по заранее заданным папкам, а обнаруживать скрытые смысловые пласты в данных. Вы объединили лингвистическую интерпретацию с математическим прогнозированием, что является основой для создания аналитических систем (мониторинг трендов, анализ соцсетей).