Федеральное государственное бюджетное образовательное учреждение

высшего образования

Московский государственный институт культуры

УТВЕРЖДЕНО
Председатель УМС
Библиотечно-информационного
факультета
Боронина Н. В.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДИСЦИПЛИНЫ (МОДУЛЯ)
Б1.В.12 Интеллектуальный анализ данных

Направление подготовки/специальности (код, наименование): 09.03.02 Информационные системы и технологии

Профиль подготовки/специализация: Информационные системы и цифровые технологии в культуре

Квалификация (степень) выпускника: Бакалавр

Форма обучения: очная

(РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов)

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели:

Целью освоения дисциплины является: сформировать у студентов комплекс знаний и практических навыков в области интеллектуального анализа текстовых данных и дата майнинга, включая работу с большими коллекциями текстов, извлечение знаний и применение ИИ технологий.

Задачи:

- изучить методы обработки и анализа больших текстовых коллекций;
- освоить технологии дата-майнинга для извлечения знаний из текстов;
- научиться применять алгоритмы машинного обучения для классификации и кластеризации текстов;
- приобрести навыки тематического моделирования и сентимент-анализа больших корпусов;
- развить способность интерпретировать результаты анализа текстовых данных для решения прикладных задач.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ВО

Дисциплина «Интеллектуальный анализ данных» входит в состав Блока 1 «Дисциплины (модули)» и относится к части, формируемой участниками образовательных отношений, ОПОП по направлению подготовки 09.03.02 Информационные системы и технологии, профиль - Информационные системы и цифровые технологии в культуре.
Дисциплина «Интеллектуальный анализ данных» изучается в седьмом, восьмом семестре. Входные знания, умения и компетенции, необходимые для изучения данного курса, формируются в процессе изучения таких дисциплин, как «Проектирование ИС», «Математика», «Вычислительные сети и системы» и «Современные информационные технологии и программное обеспечение». В результате освоения дисциплины формируются знания, умения и навыки, необходимые для изучения следующих дисциплин и прохождения практик.
Взаимосвязь курса с другими дисциплинами ООП способствует планомерному формированию необходимых компетенций и углубленной подготовке студентов к решению специальных практических профессиональных задач.

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс освоения дисциплины направлен на формирование компетенций (элементов следующих компетенций) в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки (специальности) 09.03.02 Информационные системы и технологии:

Перечень планируемых результатов обучения по дисциплине (модулю).

Компетенция (код и наименование)	**Индикаторы компетенций**	**Результаты обучения**

ПК-4 Готовность к информационно-аналитической деятельности и решению задач её автоматизации, интеллектуальному анализу данных	ПК 4.3. Работает с большими данными в области управления культурой	Знает: • основные методы обработки естественного языка (NLP) для больших данных; • технологии дата майнинга текстовых коллекций; • алгоритмы тематического моделирования (LDA, NMF) и кластеризации текстов; • инструменты для работы с большими текстовыми корпусами (Spark NLP, Gensim); • методы сентимент анализа и извлечения именованных сущностей (NER); • принципы работы нейросетевых моделей для анализа текстов (BERT, GPT).
		Умеет: • проводить предобработку больших текстовых коллекций (токенизация, лемматизация, удаление стоп слов); • применять методы дата майнинга для выявления закономерностей в текстах; • строить тематические модели и интерпретировать их результаты; • выполнять сентимент анализ отзывов и комментариев; • визуализировать результаты анализа текстовых данных; • использовать ИИ инструменты для автоматизации обработки текстов.
		Владеет: • навыками работы с библиотеками NLP (NLTK, SpaCy, Gensim, Transformers); • методами тематического моделирования и кластеризации текстов; • инструментами дата майнинга (Scikit learn, Spark NLP); • приёмами работы с большими текстовыми корпусами и API ИИ сервисов.

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ *(модуля)*

4.1 Объем дисциплины (модуля)

Объем (общая трудоемкость) дисциплины «Основы естественно-научных и общеинженерных знаний» составляет 4 з.е., 144 акад. часов, из них:
Контактных: 80 акад.ч.

СРС: 64 акад.ч.
Форма контроля: зачет, зачет с оценкой.

4.2. Структура дисциплины для очной формы обучения.

	Раздел дисциплины	Семестр	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах) в т.ч. в интерактивной форме					Формы текущего контроля успеваемости (по неделям семестра) Форма промежуточной аттестации (по семестрам)
			Лекции	Семинары/ практические	Консультации	ИКР	СРС	
1	Введение в анализ текстовых данных и дата майнинг.	7	2	5		2	8	Устные опросы на лекциях Проверка отчётов по практическим работам
2	Предобработка больших текстовых коллекций.	7	2	5		2	8	Мини-тесты по темам Проверка отчётов по практическим работам
3	Методы дата майнинга для текстов.	7	2	5		2	8	Устные опросы на лекциях Проверка отчётов по практическим работам
4	Тематическое моделирование больших корпусов.	7	2	7		4	8	Защита индивидуальных проектов (тематическая модель или дашборд с аналитикой текстов)
5	Сентимент анализ и извлечение информации.	8	2	5		2	8	Устные опросы на лекциях Проверка отчётов по практическим работам
6	Нейросетевые методы для анализа текстов.	8	2	5		2	8	Мини-тесты по темам Проверка отчётов по практическим работам
7	Дата майнинг в социальных медиа.	8	2	5		2	8	Мини-тесты по темам Проверка отчётов по практическим работам
8	Комплексный анализ текстовых данных.	8	2	7		4	8	Защита индивидуальных проектов (тематическая модель или дашборд с аналитикой текстов)

	Форма итогового контроля							Экзамен Зачет с оценкой
	Всего 144 час		**16**	**44**		**20**	**64**	

ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Текущий контроль выполнения заданий (контроль формирования компетенций) осуществляется регулярно, начиная с первой недели семестра (входящий контроль). Текущий контроль освоения отдельных разделов дисциплины осуществляется при помощи опроса, заданий (проектных и практических) и тестового материала в завершении изучения каждого раздела. Система текущего контроля успеваемости служит не только оценке уровня компетентностной подготовки обучающегося и способствует в дальнейшем наиболее качественному и объективному оцениванию его в ходе промежуточной аттестации, но и самооценке обучающегося, стимулируя его усилия.

Промежуточная аттестация проводится в форме зачета, зачета с оценкой:

- теоретический вопрос (оценка знаний методов дата майнинга и NLP);
- практическое задание (анализ большого текстового корпуса);
- защита проекта (тематическая модель коллекции текстов или дашборд с аналитикой отзывов).

СИСТЕМА ОЦЕНИВАНИЯ

Форма контроля	**Компетенция**	**Оценка**
Текущий контроль: *- опрос* *- выполнение практических работ*	*ПК-4.3.*	*зачтено/не зачтено* *зачтено/не зачтено*
Промежуточная аттестация *Экзамен* *Зачет с оценкой*	*ПК-4.3.*	*отлично/хорошо/удовлетворительно/неудовлетворительно* *зачтено (отлично, хорошо, удовлетворительно)/ не зачтено*

КРИТЕРИИ ОЦЕНКИ РЕЗУЛЬТАТОВ ПО ДИСЦИПЛИНЕ

Оценка по дисциплине	**Критерии оценки результатов обучения по дисциплине**
«отлично»/ «зачтено (отлично)»	Выставляется обучающемуся, если компетенция(ии), закрепленная за дисциплиной, сформирована (по индикаторам/ результатам обучения в формате знать-уметь-владеть) в полном объеме на уровне «высокий», и обучающийся демонстрирует как результат обучения следующие знания, умения и навыки: обучающийся глубоко и прочно усвоил теоретический и практический материал,

Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
	продемонстрировал это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет сочетать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации.
«хорошо»/ «зачтено (хорошо)»	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне «хороший».
«удовлетворительно»/ «зачтено (удовлетворительно)»	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне «достаточный».
«неудовлетворительно»/ «не зачтено» (неудовлетворительно)	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами.

Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
	Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

ТЕСТ ПО ДИСЦИПЛИНЕ:

Вариант 1

Часть 1. Закрытые вопросы (выберите один правильный ответ)

1. Какой метод используется для векторного представления текстов на основе частоты терминов?
   а) PCA;
   б) TF-IDF;
   в) K-means;
   г) LDA.
2. Какой инструмент подходит для лемматизации русских текстов?
   а) OpenCV;
   б) SpaCy;
   в) Matplotlib;
   г) TensorFlow.
3. Что такое сентимент-анализ?
   а) определение тональности текста;
   б) извлечение ключевых слов;
   в) классификация документов;
   г) построение тематических моделей.
4. Какой алгоритм применяется для тематического моделирования текстовых коллекций?
   а) SVM;
   б) LDA;
   в) CNN;
   г) KNN.
5. Какой инструмент используется для визуализации результатов тематического моделирования?
   а) pyLDAvis;
   б) BeautifulSoup;
   в) NumPy;
   г) Requests.

Часть 2. Открытые вопросы

6. Перечислите 3 этапа предобработки текстовых данных перед анализом.
7. Опишите кратко принцип работы алгоритма LDA (не более 50 слов).
8. Приведите пример практического применения сентимент-анализа в культурной сфере (1–2 предложения).
9. Назовите 2 инструмента для сбора данных из социальных сетей через API.

10. Какие метрики используются для оценки качества тематической модели? Назовите 2 метрики.

Ключи к варианту 1

Закрытые вопросы:
1 — б; 2 — б; 3 — а; 4 — б; 5 — а.

Открытые вопросы:
6. Токенизация, лемматизация/стемминг, удаление стоп-слов.
7. LDA (Latent Dirichlet Allocation) — вероятностная модель, которая представляет документы как смеси тем, а темы — как распределения по словам.
8. Анализ отзывов о выставках для определения общего настроения посетителей и выявления проблемных моментов.
9. Tweepy (для Twitter), VK API (для ВКонтакте).
10. Coherence score, perplexity.

Вариант 2

Часть 1. Закрытые вопросы (выберите один правильный ответ)

1. Какой метод преобразует слова в векторные представления с учётом семантики?
   а) One-Hot Encoding;
   б) Word2Vec;
   в) TF-IDF;
   г) PCA.
2. Что такое NER в контексте обработки текстов?
   а) метод кластеризации;
   б) извлечение именованных сущностей;
   в) алгоритм классификации;
   г) способ визуализации.
3. Какой инструмент используется для работы с трансформерами (BERT, GPT)?
   а) NLTK;
   б) Hugging Face;
   в) Gensim;
   г) OpenCV.
4. Что означает метрика coherence score?
   а) точность классификации;
   б) качество тематической модели;
   в) скорость обработки данных;
   г) размер корпуса текстов.
5. Какой метод подходит для кластеризации коротких текстов (например, твитов)?
   а) ARIMA;
   б) K-means;
   в) SVM;
   г) ANOVA.

Часть 2. Открытые вопросы

6. Назовите 3 проблемы при работе с большими текстовыми коллекциями.
7. Кратко опишите принцип работы Word2Vec (не более 50 слов).

8. Приведите пример использования тематического моделирования в анализе литературных произведений (1–2 предложения).
9. Какие инструменты можно использовать для построения дашбордов с аналитикой текстов? Назовите 2 инструмента.
10. Что такое fine-tuning в контексте нейросетевых моделей? Кратко объясните (30–50 слов).

Ключи к варианту 2

Закрытые вопросы:
1 — б; 2 — б; 3 — б; 4 — б; 5 — б.

Открытые вопросы:
6. Дублирование данных, шум в текстах, высокая размерность векторных представлений.
7. Word2Vec — метод обучения векторных представлений слов на основе их контекста: слова с похожим значением имеют близкие векторы.
8. Выявление ключевых тем в романах Достоевского для анализа эволюции его творчества.
9. Power BI, Tableau.
10. Fine-tuning — дообучение предобученной модели на специфическом датасете для повышения качества решения конкретной задачи.

Примерный перечень вопросов к экзамену:

1. Что такое интеллектуальный анализ данных (ИАД)? Охарактеризуйте его ключевые задачи и отличия от традиционного анализа данных.
2. В чём специфика применения ИАД к текстовым данным? Приведите 3–4 примера реальных бизнес-задач, решаемых с помощью ИАД текстовых данных.
3. Перечислите и кратко опишите основные этапы предобработки текстовых данных перед анализом.
4. Что такое токенизация? Опишите подходы к токенизации для языков с пробелами (английский) и без пробелов (китайский).
5. В чём разница между стеммингом и лемматизацией? Приведите примеры для русского и английского языков. Назовите популярные стеммеры и лемматизаторы.
6. Объясните принцип работы метода TF-IDF. Как рассчитывается IDF? Приведите формулу и пример расчёта для небольшого текста.
7. Что такое Word2Vec? Опишите архитектуру Skip-gram и CBOW, укажите их сильные и слабые стороны.
8. В чём отличие Doc2Vec от Word2Vec? Какие существуют варианты реализации Doc2Vec (PV-DM, PV-DBOW)?
9. Сравните TF-IDF, Word2Vec и Doc2Vec по следующим критериям: интерпретируемость, размерность вектора, учёт контекста, вычислительная сложность. Представьте результат в виде таблицы.
10. Опишите алгоритм K-means для кластеризации текстов. Как выбрать оптимальное число кластеров (метод локтя, silhouette score)?
11. Что такое иерархическая кластеризация? Постройте дендрограмму для небольшого набора текстов и объясните, как по ней определить кластеры.
12. Что такое тематическое моделирование? В чём суть вероятностной модели LDA? Какие гиперпараметры есть у LDA (α, β) и как они влияют на результат?
13. Как интерпретировать результаты LDA? Приведите пример темы (топ-слов) и объясните, как понять, что тема осмысленная.
14. Назовите метрики для оценки качества тематических моделей (perplexity, coherence score). В чём их преимущества и недостатки?

15. Как визуально интерпретировать результаты тематического моделирования? Опишите, какие паттерны в pyLDAvis говорят о хорошей/плохой модели.
16. Перечислите основные источники текстовых данных для анализа. Укажите плюсы и минусы каждого (веб-страницы, соцсети, архивы, оцифрованные книги).
17. Какие юридические и этические ограничения нужно учитывать при сборе текстовых данных из соцсетей и веб-страниц?
18. Какие проблемы возникают при работе с большими текстовыми коллекциями (миллионы документов)? Опишите 2–3 способа их решения (распараллеливание, выборка, инкрементальное обучение).
19. Сравните NLTK и SpaCy для предобработки текстов: скорость, функциональность, поддержка языков, удобство API. Приведите пример кода токенизации на обоих инструментах.
20. Опишите функционал pyLDAvis. Как с его помощью оценить качество тематической модели и объяснить результаты не-эксперту?

Примерный перечень вопросов для к зачету с оценкой:

1. Сентимент-анализ: методы определения тональности текстов, метрики оценки качества.
2. Извлечение именованных сущностей (NER): задачи, алгоритмы, применение в культурной сфере.
3. Нейросетевые методы для анализа текстов: трансформеры (BERT, RoBERTa), их преимущества перед традиционными методами.
4. Fine-tuning предобученных языковых моделей: назначение и процесс.
5. Анализ социальных медиа: сбор данных через API, анализ хештегов и трендов.
6. Дата-майнинг в социальных сетях: выявление влиятельных пользователей и сообществ.
7. Комплексный анализ текстовых данных: интеграция методов дата-майнинга и машинного обучения.
8. Построение дашбордов с аналитикой текстов: ключевые показатели и визуализация.
9. Этические и правовые аспекты анализа текстовых данных (ФЗ-152, ФЗ-149).
10. Применение ИИ-инструментов (Yandex GPT и др.) для автоматизации анализа текстов: возможности и ограничения.

Примеры практико-ориентированных задач к зачёту с оценкой

Задача 1. Сентимент-анализ

Дан набор из 5 коротких отзывов о библиотеке.

1. Проведите ручную разметку тональности каждого отзыва: *положительный*, *отрицательный*, *нейтральный*.
2. Для одного из отзывов приведите 2–3 лексических маркера (слова/фразы), по которым вы определили тональность.
3. Назовите две метрики качества для оценки модели сентимент-анализа и кратко поясните, что они измеряют.

Пример отзыва: «Библиотекари скучные, но обстановка красивая».

Задача 2. Извлечение именованных сущностей (NER)

Дан текст на русском языке о культурном событии (фестиваль, выставка и т.д.).

1. Выделите и классифицируйте именованные сущности по типам: *PER* (персона), *ORG* (организация), *LOC* (место), *DATE* (дата).
2. Кратко опишите, как результаты NER могут помочь в анализе культурной сферы (приведите 1–2 примера применения).

Пример текста: «В Третьяковской галерее 15 мая открылась выставка работ Ильи Репина».

Задача 3. Нейросетевые методы для анализа текстов

Сравните традиционный метод TF-IDF + классификатор (например, SVM) с моделью BERT для задачи классификации новостей по темам.

1. В таблице из 3 строк укажите:
   - **Учёт контекста** (да/нет, кратко поясните);
   - **Обработка омонимии** (как решается);
   - **Требования к объёму данных** (большой/малый).

Задача 4. Fine-tuning предобученных моделей

Опишите пошаговый процесс fine-tuning модели RuBERT для классификации отзывов о музеях (классы: *положительный*, *отрицательный*).

1. Перечислите 4 ключевых шага (от подготовки данных до вывода).
2. Укажите, какие слои модели обычно замораживают/размораживают на этапе fine-tuning и почему.

Задача 5. Анализ социальных медиа

Вы собираете данные о хештегах #МоскваБиблиотека через API VK за последнюю неделю.

1. Составьте список из 5 метрик для анализа трендов по этому хештегу (например, *количество постов в день*).
2. Предложите способ визуализации этих метрик (тип графика + ось X/Y).

Задача 6. Дата-майнинг в соцсетях

Дано: граф взаимодействий пользователей в соцсети (вершины — пользователи, рёбра — подписки/лайки).

1. Назовите два алгоритма для выявления влиятельных пользователей (лидеров мнений) и кратко опишите принцип их работы (1–2 предложения на алгоритм).
2. Приведите пример метрики влиятельности (например, PageRank) и объясните, что она измеряет.

Задача 7. Комплексный анализ текстовых данных

Опишите последовательность этапов для анализа отзывов о мобильном приложении, объединяющий методы дата-майнинга и машинного обучения.

1. Перечислите 3 этапа обработки (от сырых текстов до выводов).
2. Для каждого этапа укажите:
    - метод дата-майнинга/ML;
    - цель этапа (что он даёт для общего анализа).

Задача 8. Построение дашбордов с аналитикой текстов

Вам нужно создать дашборд для мониторинга тональности упоминаний бренда в соцсетях.

1. Выберите 4 ключевых показателя (KPIs) для отображения (например, *доля положительных упоминаний*).
2. Для каждого KPI предложите тип визуализации (столбчатая диаграмма, линейный график, круговая диаграмма и т.д.) и кратко обоснуйте выбор.

Задача 9. Этические и правовые аспекты

Вы выявляет случаи нелегитимного использования личных сообщений пользователей в мессенджере для исследования общественного мнения.

1. Укажите два положения ФЗ 152-ФЗ («О персональных данных») и два положения ФЗ 149-ФЗ («Об информации…»), которые необходимо учесть.
2. Укажите два способа анонимизации данных, которые используют недобросоветстные исследователи перед анализом, чтобы снизить риски нарушения закона.

Задача 10. Применение ИИ-инструментов

Используйте Yandex GPT для генерации краткого резюме (3–4 предложения) по тексту новости (предоставлен экзаменатором).

1. Скопируйте запрос к модели и получившееся резюме.
2. Оцените качество результата по двум критериям: *точность фактов*, *полнота*. Приведите один пример ошибки/упущения, если они есть.
3. Кратко (1–2 предложения) укажите, в каких задачах анализа текстов такие инструменты наиболее полезны, а в каких — ненадёжны.

ОЦЕНОЧНЫЕ СРЕДСТВА ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ

Практические задания:

Задание 1. Литературный анализ

- Задача: провести тематическое моделирование сборника стихов русских поэтов Серебряного века.
- Данные: корпус текстов стихотворений (Блок, Ахматова, Маяковский и др.).
- Инструменты: Python, Gensim (LDA), NLTK/SpaCy.
- Результат: визуализация тем и ключевых слов для каждой темы, сравнение тематических предпочтений поэтов.

Задание 2. Политический анализ

- Задача: проанализировать тональность твитов о культурной политике России за последний год.
- Данные: твиты с хештегом #культурнаяполитика (через Twitter API).
- Инструменты: Python, TextBlob/VADER (сентимент-анализ), Matplotlib (визуализация).
- Результат: график изменения тональности обсуждений во времени, облако тегов наиболее обсуждаемых тем.

Задание 3. Анализ культурной политики

- Задача: сравнить ключевые темы в государственных программах поддержки культуры 2010 и 2020 годов.
- Данные: тексты программ с официальных сайтов Минкультуры.
- Инструменты: Python, TF-IDF, Word2Vec.
- Результат: сравнительная таблица ключевых тем, визуализация различий в приоритетах.

Задание 4. Анализ читательских предпочтений

- Задача: выявить популярные жанры в региональной библиотеке за последние 5 лет.
- Данные: статистика выдачи книг (жанр, автор, год издания).
- Инструменты: Power BI, Pandas (агрегация данных).
- Результат: дашборд с динамикой популярности жанров, топ-10 авторов по годам.

Задание 5. Мониторинг культурного разнообразия

- Задача: проанализировать репертуар театров Москвы на представленность современных российских авторов.
- Данные: афиши театров за сезон (парсинг с сайтов).
- Инструменты: Python (BeautifulSoup), SpaCy (NER для извлечения имён авторов).
- Результат: процент пьес современных российских авторов в репертуаре, топ-5.

Практические задания (на выбор):

1. Провести предобработку корпуса отзывов о культурных мероприятиях (удалить стоп-слова, привести к нижнему регистру, выполнить лемматизацию).
2. Построить TF-IDF-матрицу для небольшого корпуса текстов (5–10 документов).
3. Выполнить кластеризацию коротких текстов (например, твитов о музеях) с помощью алгоритма K-means.
4. Построить простую тематическую модель (LDA) для небольшого набора статей о культуре (3–5 тем).
5. Визуализировать результаты тематического моделирования с помощью облака слов для каждой темы.
6. Извлечь ключевые слова из текста с помощью TF-IDF.
7. Сравнить два текста на схожесть с использованием векторных представлений (косинусное сходство).
8. Подготовить отчёт с описанием этапов предобработки и анализа текстового корпуса (500–700 слов).

9. Выполнить сентимент-анализ корпуса отзывов о выставках (определить общую тональность, построить график распределения положительных/отрицательных отзывов).
10. Извлечь именованные сущности (люди, места, организации) из статей о культурных событиях с помощью SpaCy.
11. Провести тематическое моделирование большого корпуса статей о культурной политике (100+ документов, 5–7 тем) с использованием LDA или NMF.
12. Классифицировать тексты (например, новости о театре, музыке, изобразительном искусстве) с помощью модели BERT (Hugging Face).
13. Собрать данные из соцсетей (например, твиты с хештегом #культурнаяполитика) через API и выполнить анализ трендов (частота хештегов, ключевые слова).
14. Построить дашборд в Power BI или Tableau с визуализацией результатов анализа текстов (тональность, темы, ключевые слова, динамика обсуждений).
15. Выполнить fine-tuning модели BERT на корпусе текстов о культуре для задачи классификации (например, «актуальные события» vs «архивные материалы»).
16. Разработать сценарий использования ИИ-ассистента (на базе Yandex GPT) для анализа запросов посетителей культурного учреждения и генерации отчётов.
17. Оценить качество модели тематического моделирования (LDA) с использованием метрик coherence score и perplexity.
18. Подготовить итоговый отчёт (800–1000 слов) с описанием комплексного анализа текстового корпуса: этапы, методы, результаты, выводы и рекомендации.