

**Федеральное государственное образовательное бюджетное
учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)**

**Департамент анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных**

УТВЕРЖДАЮ

Проректор по учебной
и методической работе

_____ Е.А. Каменева
25.04.2023 г.

Макрушин С.В., Блохин Н.В.

Технологии обработки больших данных

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
01.03.02 - Прикладная математика и информатика,
ОП "Прикладное машинное обучение"

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол №31 от 18.04.2023г.)*

*Одобрено Советом учебно-научного
Департамента анализа данных и машинного обучения
(протокол №2 от 29.03.2023г.)*

Москва 2023

Содержание

1. Наименование дисциплины.....	2
2.Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	2
3. Место дисциплины в структуре образовательной программы	3
4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся.....	3
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	3
5.1. Содержание дисциплины	3
5.2. Учебно-тематический план.....	6
5.3. Содержание семинаров, практических занятий	7
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	8
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы.....	8
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю.....	9
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине.....	13
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	17
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	18
10. Методические указания для обучающихся по освоению дисциплины.	19
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем.....	20
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.	21

1. Наименование дисциплины

«Технологии обработки больших данных».

2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКН-4	Способен проектировать и реализовывать прикладные программные системы в соответствии с анализом задачи и требований к ним	Демонстрирует базовые знания о существующих математических методах и системах программирования.	Знать программные системы, позволяющие обрабатывать большие данные. Уметь использовать программные системы, позволяющие обрабатывать большие данные.
		Использует и адаптирует существующие математические методы и системы программирования для решения прикладных задач.	Знать принципы построения прикладных систем на основе инфраструктурных решений для обработки больших данных. Уметь реализовывать прикладные системы на основе инфраструктурных решений для обработки больших данных.
		Владеет навыками проектирования и разработки компонентов программного обеспечения на основе современных парадигм, технологий и языков программирования.	Знать принципы разработки программных системы для обработки больших данных. Уметь разрабатывать программные системы для обработки больших данных.
		Применяет методы машинного обучения для решения прикладных задач анализа данных.	Знать технологии обработки больших данных применимые для решения задач машинного обучения. Уметь использовать технологии обработки больших данных применимые для решения задач машинного обучения.

3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки больших данных» является дисциплиной Общефакультетского (предпрофильного) цикла по направлению подготовки 01.03.02 - Прикладная математика и информатика, ОП "Прикладное машинное обучение".

4. Объем дисциплины(модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 5 (в часах)
Общая трудоёмкость дисциплины	5/180	180
<i>Контактная работа- Аудиторные занятия</i>	68	68
Лекции	16	16
Семинары, практические занятия	52	52
<i>Самостоятельная работа</i>	112	112
Вид текущего контроля	Экзамен	Экзамен
Вид промежуточной аттестации	Проектная работа	Проектная работа

5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема 1. Параллельная обработка данных.

В рамках темы рассматривается специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение.

Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. Специфика различия между потоками и процессами.

Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool.

Тема 2. Библиотека Dask.

В рамках темы рассматривается библиотека для анализа больших объемов данных Python Dask, различные предлагаемые ей подходы к обработке данных. Рассматривается структура данных Dask.Array, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Array операции и ее отличия от NumPy ndarray.

Тема 3. Подход к обработке данных в стиле Map-Reduce.

Рассматривается организация вычислений с помощью подхода Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag. Рассматривается структура данных Dask.Bag, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Bag операции.

Тема 4. Взаимодействие с базой данных в приложениях обработки данных.

В рамках темы рассматривается взаимодействие из Python с базой данных. Решение рассматривается на примере API SQLite. Рассматриваются базовые возможности работы с транзакциями.

Тема 5. Обработка табличных данных с помощью Dask.

Рассматривается структура данных `Dask.DataFrame`, специфика ее реализации и применения, процедура создания, ограничения использования `Dask.DataFrame`. Рассматриваются операции мэппинга в `Dask.DataFrame` и операции `Dask.DataFrame` работающие со скользящим окном.

Тема 6. Хранилища In-Memory для различных приложений.

Рассматриваются общие принципы In-Memory хранилищ и решений, предлагающих работу с большими массивами данных в памяти. Рассматриваются возможности и специфика использования In-Memory хранилища `Redis`. Рассматриваются приложения для использования библиотеки `Faiss`, функциональные возможности данного решения.

Тема 7. Распараллеливание выполнения алгоритмов с помощью Dask.

Инструменты, предназначенные для обеспечения параллельных вычислений и обработки данных в Python `Dask` и модуль `dask.delayed`. `Dask.delayed` - это удобная библиотека, которая упрощает написание параллельного кода. `Dask` предоставляет мощные API для управления несколькими ядрами, кластерами и графическими процессорами. Он может масштабироваться до тысяч ядер и использоваться для параллельной обработки данных. `Dask.delayed` - это асинхронный планировщик задач, который позволяет вам вызывать функции асинхронно и возвращает результаты по мере их поступления, что позволяет строить сложные конвейеры обработки данных. Организация эффективных и воспроизводимых структур параллельной обработки данных с помощью `Dask.delayed`.

Тема 8. Решения Hadoop и Spark для обработки больших данных.

Обзор решений Hadoop и Spark для обработки больших данных. Это включает в себя рассмотрение инструментов Apache Hadoop и Apache Spark и их применение к решению практических проблем обработки больших данных. Рассматриваются основные блоки решений Hadoop, Spark, в том числе хранилище данных Hadoop Distributed File System (HDFS), MapReduce framework и Apache Spark. Рассматриваются практические примеры для использования решений Spark / Hadoop.

5.2. Учебно-тематический план

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости
		Всего	Контактная работа- Аудиторная работа			Самосто- ятельная работа	
			Общая, в т.ч.:	Лекции	Семинары, практиче- ские занятия		
1	Параллельная об- работка данных	24	10	2	8	14	Участие в решении задач на практиче- ских заня- тиях. Об- суждения по резуль- татам са- мостоя- тельной ра- боты
2	Библиотека Dask	22	8	2	6	14	
3	Подход к обра- ботке данных в стиле Map-Reduce	22	8	2	6	14	
4	Взаимодействие с базой данных в приложениях об- работки данных	20	6	2	4	14	
5	Обработка таб- личных данных с помощью Dask	22	8	2	6	14	
6	Хранилища In- Мемогу для раз- личных приложе- ний	24	10	2	8	14	Участие в решении задач на практиче- ских заня- тиях. Об- суждения
7	Распараллелива- ние выполнения	22	8	2	6	14	

	алгоритмов с помощью Dask						по результатам самостоятельной работы
8	Решения Hadoop и Spark для обработки больших данных	24	10	2	8	14	
	В целом по дисциплине	180	68	16	52	112	Согласно учебному плану: проектная работа
	Итого в %		38	24	76	62	

5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8,9 (указывается раздел и порядковый номер источника)	Формы проведения занятий
Параллельная обработка данных	• специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. • Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. • Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. • Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool. 8[1], 8[2], 9[5], 9[7]	Интерактивная форма, работа на компьютере
Библиотека Dask	• Подход к обработке данных с помощью библиотеки Dask. • Структура данных Dask.Array – принцип работы, API, примеры использования. 8[1], 8[2], 9[8], 9[10], 9[11]	Интерактивная форма, работа на компьютере
Подход к обработке данных в стиле Map-Reduce	• Решение задач обработки данных с помощью подхода Map-Reduce • Структура данных Dask.Bag – принцип работы, API, примеры использования. 8[1], 8[2], 9[8], 9[10], 9[11]	Интерактивная форма, работа на компьютере
Взаимодействие с базой данных в приложениях обработки данных.	• Взаимодействие из Python с базой данных с помощью API SQLite. 8[1], 8[2]	Интерактивная форма, работа на компьютере
Обработка таблич-	• Структура данных Dask.DataFrame – принцип работы, API, примеры использования.	Интерактивная форма, работа

ных данных с помощью Dask	8[1], 8[2], 9[8], 9[10], 9[11]	на компьютере
Хранилища In-Memory для различных приложений	- Базовые сценарии использования Redis. - Принципы использования Faiss. 8[1], 8[2]	Интерактивная форма, работа на компьютере
Распараллеливание выполнения алгоритмов с помощью Dask	Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool. 8[1], 8[2], 9[5], 9[7]	Интерактивная форма, работа на компьютере
Решения Hadoop и Spark для обработки больших данных	- Знакомство с использованием Hadoop. - Знакомство с использованием Spark. - Решение базовых задач обработки данных с помощью Hadoop/Spark.	Интерактивная форма, работа на компьютере

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
Параллельная обработка данных	- Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. - Специфика различия между потоками и процессами. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Библиотека Dask	Настройка Dask для работы на нескольких компьютерах с помощью Dask.Distributed	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Подход к обработке данных в стиле Map-Reduce	Изучение API Dask.Bag для обработки строковых данных.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Взаимодействие с базой данных в приложениях обработки	Базовые возможности работы с транзакциями с помощью API SQLite.	Обзор литературы и веб-источников. Самостоя-

данных.		тельное освоение инструментов аналитической обработки. Решение задач.
Обработка табличных данных с помощью Dask	Рассмотрение сценариев последовательного использования Dask.Bag и Dask.DataFrame	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Хранилища In-Memory для различных приложений	Реализация поиска ближайших значений для эмбедингов строковых данных с помощью Faiss.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Распараллеливание выполнения алгоритмов с помощью Dask	Организация вычислений с помощью API Dask Delayed.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Решения Hadoop и Spark для обработки больших данных	Разворачивание Apache Spark с помощью контейнера docker.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерные вопросы к проектной работе

1. Большие данные – определение и причины возникновения задач обработки больших данных
2. Специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений
3. Выбор типичных средств обработки данных, адекватных различным объемам данных; принцип обработки данных на базе операций map / filter / reduce
4. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение

5. Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения
6. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем
7. Различия между потоками и процессами, различие между различными планировщиками в Dask
8. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
9. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
10. Dask.Array – структура данных, специфика реализации и применения, процедура создания
11. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
12. Распараллеливание алгоритмов с помощью `dask.delayed` – принцип и примеры использования
13. Дополнительные параметры декоратора `dask.delayed` – назначение и примеры использования
14. Использование `dask.delayed` для объектов и операции над объектами `dask.delayed`, включая ограничения их использования
15. Dask.DataFrame – структура данных, специфика реализации и применения, процедура создания Dask.DataFrame
16. Ограничения использования Dask.DataFrame и операции мэппинга в Dask.DataFrame
17. Поддержка Dask.DataFrame операций работающих со скользящим окном
18. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования
19. Dask.Bag – структура данных, специфика реализации и применения, процедура создания DaskBag

20. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в Dask.Bag

21. API Dask.Bag – функции мэппинга, фильтрации и преобразования

Примерные задания проектной работы

Задание 1

1. В массиве чисел, хранящихся в файле `finance.hdf5`, найти строку (вывести ее индекс и содержащиеся значения), в которой более всего значений, превышающих среднее значение по всему массиву. Для расчётов использовать `dask.array`.
2. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество строк, в которых более 600 значений больше среднего значения по всему массиву. Для расчётов использовать `dask.array`.
3. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество значений, не отклоняющихся от среднего значения более чем на 3 стандартных отклонения. Для расчетов использовать `dask.array`

Задание 2

1. В `accounts/*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений, кратных трем. Выполнить задание с использованием Dask, распараллелив процесс обработки данных
2. В `accounts/*.csv` найти `id`, для которого сумма положительных значений в столбце `amount` наибольшая. Выполнить задание с использованием Dask, распараллелив процесс обработки данных.
3. В `accounts/*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений между 1000 и 1500. Выполнить задание с использованием Dask, распараллелив процесс обработки данных

Задание 3

Датасет: all_k.zip

Подсчитать, сколько раз в текстовых файлах, лежащих в all_k.zip, встречаются предложения трех видов: вопросительные (в окончании имеют вопросительный знак), побудительные (в окончании имеют восклицательный знак и не имеют вопросительного) и повествовательные (в окончании имеют точку или троеточие, при этом нужно исключить учет точек, встречающихся в сокращениях, таких как "т.к.").

Выполнить задание с использованием Dask (корректным!), распараллелив процесс обработки данных (использование Dask должно приводить к истинной параллельной обработке данных).

Задание 4

Датасет: all_k.zip

Подсчитать, сколько раз встречается каждое из личных местоимений в именительном падеже (полный список: я, ты, он, она, оно, мы, вы, они) в текстовых файлах, лежащих в папке: all_k.zip.

Выполнить задание с корректным использованием Dask, распараллелив процесс обработки данных (использование Dask должно приводить к истинной параллельной обработке данных).

Примерная тематика проектной работы

1. Прогнозная аналитика и моделирование объемов продаж акций
2. Визуализация аналитических данных в области макроэкономики
3. Визуализация аналитических данных Московской биржи
4. Использование технологии больших данных для анализа портфельных рисков
5. Использование параллельных вычислений реализации численных методов решения математических задач

6. Анализ и сравнение различных фреймворков для визуализации данных
7. Применение распределенных вычислений и экосистемы Hadoop для решения задачи анализа данных
8. Анализ больших данных для построения прогнозов на рынке ценных бумаг
9. Использование больших данных для оценки кредитоспособности контрагентов на основе анализа текстов новостей
10. Проведение анализа собранных из внешних источников данных

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных.

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 2. **«Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине».**

Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний

Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания

ПКН-4 Способен проектировать и реализовывать прикладные программные системы в соответствии с анализом задачи и требований к ним	Демонстрирует базовые знания о существующих математических методах и системах программирования.	Знать программные системы, позволяющие обрабатывать большие данные. Уметь использовать программные системы, позволяющие обрабатывать большие данные.	Перечислить ключевые библиотеки / фреймворки для обработки больших данных. Предложить программный стек решения задачи подсчета количества слов для корпуса текстов в 100 Мб, 1 Гб, 10 Гб, 1 Тб и 50 Тб.
	Использует и адаптирует существующие математические методы и системы программирования для решения прикладных задач.	Знать принципы построения прикладных систем на основе инфраструктурных решений для обработки больших данных. Уметь реализовывать прикладные системы на основе инфраструктурных решений для обработки больших данных.	Предложить архитектуру решения задачи подсчета количества слов для корпуса текстов в 100 Мб, 1 Гб, 10 Гб, 1 Тб и 50 Тб. Создать программное решение задачи подсчета количества слов для корпуса текстов в 100 Мб, 1 Гб.
	Владеет навыками проектирования и разработки компонентов программного обеспечения на основе современных парадигм, технологий и языков программирования.	Знать принципы разработки программных системы для обработки больших данных. Уметь разрабатывать программные системы для обработки больших данных.	Спроектировать программное решение задачи подсчета количества слов для корпуса текстов в 100 Мб, 1 Гб, 10 Гб, 1 Тб и 50 Тб. Реализовать программное решение задачи подсчета количества слов для корпуса текстов в 100 Мб, 1 Гб.
	Применяет методы машинного обучения для решения прикладных задач анализа данных.	Знать технологии обработки больших данных применимые для решения задач машинного обучения. Уметь использовать технологии обработки больших данных применимые для решения задач машинного обучения.	Предложить решения для поиска перефразированных дубликатов в большом корпусе текстов на естественном языке. Создать программную реализацию решения для поиска перефразированных дубликатов в большом корпусе текстов на естественном языке.

Примерные вопросы для подготовки к экзамену

1. Большие данные – определение и причины возникновения задач обработки больших данных
2. Специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений
3. Выбор типичных средств обработки данных, адекватных различным объемам данных; принцип обработки данных на базе операций map / filter / reduce
4. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение
5. Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения
6. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем
7. Профилирование реализации алгоритмов на Python, принципы решения задачи оптимизации производительности алгоритма
8. Проблема Global Interpreter Lock в Python и способы обхода ее ограничений
9. Модуль multiprocessing – назначение и основные возможности, API multiprocessing.Pool
10. Различия между потоками и процессами, различие между различными планировщиками в Dask
11. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
12. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач

13. Dask.Array – структура данных, специфика реализации и применения, процедура создания
14. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
15. Распараллеливание алгоритмов с помощью dask.delayed – принцип и примеры использования
16. Дополнительные параметры декоратора dask.delayed – назначение и примеры использования
17. Использование dask.delayed для объектов и операции над объектами dask.delayed, включая ограничения их использования
18. Dask.DataFrame - структура данных, специфика реализации и применения, процедура создания Dask.DataFrame
19. Ограничения использования Dask.DataFrame и операции мэппинга в Dask.DataFrame
20. Поддержка Dask.DataFrame операций работающих со скользящим окном
21. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования
22. Dask.Bag - структура данных, специфика реализации и применения, процедура создания DaskBag
23. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в Dask.Bag
24. API Dask.Bag – функции мэппинга, фильтрации и преобразования
25. API Dask.Bag – функции группировки и свертки
26. Принципы работы Apache Hadoop.
27. Принципы работы Apache Spark.
28. Сценарии использования Faiss.
29. Сценарии использования Redis.

Пример экзаменационного билета

1. **(20 баллов)** Большие данные – определение и причины возникновения задач обработки больших данных.

2. **(20 баллов)** Датасет: Chinook_Sqlite.sqlite

С помощью кода на Python с использованием sqlite3 и SQL решить задачу. Реализовать функции на Python:

1. Которая возвращает все имеющиеся плейлисты.
2. Которая по имени плейлиста возвращает количество треков в нем и их суммарную продолжительность.

3. **(20 баллов)** Датасет: all_k.zip

Подсчитать, сколько раз во всех текстовых файлах, лежащих в all_k.zip, встречаются реплики прямой речи, оформленные в виде диалога (В этом случае каждая реплика начинается с новой строки, перед репликами ставится тире (перед тире возможны различные пробельные символы)). Выполнить задание с использованием Dask, распараллелив процесс обработки данных.

Выполнить задание с использованием Dask (корректным!), распараллелив процесс обработки данных (использование Dask должно приводить к истинной параллельной обработке данных).

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

Основная литература:

1. Колдаев, В. Д. Структуры и алгоритмы обработки данных : учебное пособие / В. Д. Колдаев. - Москва : РИОР : ИНФРА-М, 2021. - 296 с. - ЭБС ZNANIUM.com. - URL: <https://znanium.com/catalog/product/1230215> (дата обращения: 10.05.2023). – Текст : электронный.

Дополнительная литература:

2. Нагаева, И. А. Основы алгоритмизации и программирования: практикум : учебное пособие / И. А. Нагаева, И. А. Кузнецов. – Москва : Берлин : Директ-Медиа, 2021. – 169 с. – ЭБС Университетская библиотека ONLINE. – URL: <https://biblioclub.ru/index.php?page=book&id=598404> (дата обращения: 10.05.2023). – Текст : электронный.

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Электронная библиотека Финансового университета (ЭБ)
<http://elib.fa.ru/>

2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>

3. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>

4. Электронно-библиотечная система Znanium
<http://www.znanium.com>

5. Электронно-библиотечная система издательства «ЮРАЙТ»
<https://urait.ru/>

6. Электронно-библиотечная система издательства Проспект
<http://ebs.prospekt.org/books>

7. Электронно-библиотечная система издательства Лань
<https://e.lanbook.com/>

8. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>

9. Электронная библиотека Издательского дома «Гребенников»
<https://grebennikon.ru/>

10. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>

11. Национальная электронная библиотека <http://нэб.рф/>

12. Pyru 1.0.9 [Электронный ресурс]: сайт. – Режим доступа:
<https://pypi.python.org/pypi/pylru>

13. Python Data Analysis Library [Электронный ресурс]: сайт. – Режим доступа: <http://pandas.pydata.org/>
14. Python Documentation [Электронный ресурс]: сайт. – Режим доступа: <http://python.org/doc/>
15. Python Standard Library [Электронный ресурс]: сайт. – Режим доступа: <https://docs.python.org/2/library/>
16. Scikit-learn Machine Learning in Python [Электронный ресурс]: сайт. – Режим доступа: <http://scikit-learn.org>
17. Официальный сайт продукта <https://www.python.org/>
18. Каталог курсов Интернет Университета Информационных Технологий <http://www.intuit.ru/>
19. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
20. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
21. Pandas User Guide <http://pandas.pydata.org/pandas-docs/stable/>
22. Dask User Guide <https://docs.dask.org/en/latest/>
23. Dask User Guide <https://docs.dask.org/en/latest/>
24. Matplotlib User Guide // <https://matplotlib.org/stable/users/index.html>
25. Seaborn User Guide // <https://seaborn.pydata.org/tutorial.html>

10. Методические указания для обучающихся по освоению дисциплины

При изучении теоретического материала необходимо опираться на рабочую программу дисциплины, материалы лекций и литературу из основного списка. Кроме этого, необходимо активно работать с Интернет-источниками и пособиями других авторов, помогающими усвоить материал отдельных разделов программы.

Необходимо конспектировать лекции, помечая сложные и непонятные моменты с тем, чтобы задать вопросы лектору в конце лекции или же на консультации.

При подготовке к семинарским занятиям необходимо изучить вопросы, вынесенные на самостоятельное изучение, так как семинарские занятия предполагают их обсуждение и дискуссию по теме; кроме того, задания для самостоятельной работы необходимы для того, чтобы успешно выполнить самостоятельные задания на семинарах.

Индивидуальные задания для работы на компьютере, файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем

11.1. Комплект лицензионного программного обеспечения:

1. Пакет офисных программ
2. Антивирус Kaspersky

11.2. Современные профессиональные базы данных и информационные справочные системы

1. Информационно-правовая система «Гарант»
2. Информационно-правовая система «Консультант Плюс»
3. Электронная энциклопедия: <http://ru.wikipedia.org/wiki/Wiki>
4. Система комплексного раскрытия информации «СКРИН» - <http://www.skrin.ru/>

11.3. Сертифицированные программные и аппаратные средства защиты информации

- не используются

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Для проведения лекций и практических занятий необходима аудитория, оснащенная проектором и компьютерами с постоянным подключением к сети Интернет.