Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

УТВЕРЖДАЮ
декан факультета вычислительной
математики и кибернетики

_____ / И.А. Соколов /
«27» сентября 2023г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине

Методы и системы обработки больших данных

Уровень высшего образования:

бакалавриат

Направление подготовки / специальность:

02.03.02 "Фундаментальная информатика и информационные технологии" (3++)

Направленность (профиль) ОПОП:

Искусственный интеллект и анализ данных

Форма обучения:

очная

Рассмотрен и утвержден
на заседании Ученого совета факультета ВМК
(протокол №7, от 27 сентября 2023 года)

Москва 2023

1. ФОРМЫ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

В процессе и по завершении изучения дисциплины оценивается формирование у студентов следующих компетенций:

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
ПК-8. Способен разрабатывать системы анализа больших данных	ПК-8.1. Разрабатывает программные компоненты извлечения, хранения, подготовки больших данных с учетом вариантов использования больших данных, определений, словарей и эталонной архитектуры больших данных ПК-8.2. Разрабатывает программные компоненты обработки, удаленной, распределенной и объединенной аналитики, использования результатов анализа, описания и управления качеством и достоверностью больших данных	ПК-8.1. 3-1. Знает общедоступные репозитории и специализированные библиотеки, содержащие наборы больших данных ПК-8.1. 3-2. Знает принципы работы экосистемы Hadoop, фреймворка SPARK ПК-8.1. 3-3. Знает устройство интерфейсов между реляционными SQL-хранилищами данных и нереляционными NoSQL-хранилищами данных ПК-8.1. 3-4. Знает предметно-ориентированные языки ПК-8.1. У-1. Умеет настраивать и оптимизировать конфигурацию программного и аппаратного обеспечения с целью интеграции больших данных ПК-8.1. У-2. Умеет разрабатывать программное обеспечение для очистки и валидации наборов больших данных ПК-8.1. У-3. Умеет выполнять потоковую обработку данных (data streaming, event processing) ПК-8.1. У-4. Умеет использовать шины данных (Apache Kafka) ПК-8.1. У-5. Умеет использовать языки запросов, в том числе нереляционных, для поддержки различных типов данных (например, XML, RDF, JSON, мультимедиа) и

		операций с большими данными (например, матричные операции) ПК-8.2. 3-1. Знает принципы и методы анализа больших данных, включая спецификации и стандартизацию метаданных ПК-8.2. 3-2. Знает устройство и принципы работы систем обработки и анализа больших массивов данных (SQL. NoSQL, Hadoop, ETL) ПК-8.2. 3-3. Знает архитектуру и принципы работы промышленных решений, созданных на основе искусственного интеллекта ПК-8.2. 3-4. Знает методы и технологии машинного обучения на больших данных ПК-8.2. У-1. Умеет разрабатывать программное обеспечение для анализа больших данных. ПК-8.2. У-2. Умеет разрабатывать программные и технические средства визуализации больших данных и результатов их анализа. ПК-8.2. У-3. Умеет использовать системы обработки и анализа больших массивов данных (SQL, NoSQL, Hadoop, ETL процессы и инструменты) ПК-8.2. У-4. Умеет использовать технологии науки о данных и больших данных в разработке для решения практических задач промышленности ПК-8.2. У-5. Умеет описывать и управлять качеством и достоверностью больших данных

1.1. Текущий контроль успеваемости

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

Практические занятия

1. Работа с изображениями на платформе Python.
2. Детектирование объектов и сегментация изображения с использованием библиотек Python
3. Распознавание лиц с помощью MTCNN
4. Распознавание эмоций
5. Рекуррентные нейронные сети
6. Работа с графами в Neo4J

Примеры тестовых заданий

Вопрос №1. Укажите свойства и функции, которыми **не** обладает движок обработки данных Apache Spark

1. поддержка языков программирования Java, Scala, Python
2. высокое быстродействие (по сравнению с Hadoop MapReduce)
3. сохранение на диск всех промежуточных результатов для возможности восстановления после сбоя
4. небольшие требования к процессору и объему оперативной памяти

Вопрос №2. Какие операции Apache Spark в общем случае приведут к полному перераспределению (shuffle) данных по узлам кластера

1. map()
2. repartition()
3. coalesce()
4. sortBy()

Вопрос №3. Укажите сценарии, для которых применим Apache Spark

1. Чтение данных из реляционной базы данных, конвертация в Avro-формат, сжатие в GZIP и сохранение на HDFS
2. Соединение (JOIN) данных из таблицы в Oracle Datavse и файла в формате Excel
3. Запуск задач обработки данных по расписанию
4. Тестирование данных на обученной модели
5. Буферизация потока данных для равномерной загрузки кластера

Вопрос №4. Выберите этапы обработки данных, которые могут быть применимы до обучения модели:

1. Преобразование в табличный формат
2. Удаление некорректных данных
3. Замена слов на синонимы
4. Визуализация данных в виде графика

Вопрос №5. Выберите возможные методы нормализации при анализе текста на естественном языке:

1. Стемминг
2. TF-IDF
3. Удаление stop-слов
4. Замена слов на синонимы

Вопрос №6. Какие из следующих программных продуктов предназначены для машинного обучения на больших данных

1. Apache Hadoop
2. Apache SparkML

3. Skikit-learn
4. H2O
5. Tensorflow

Вопрос №7. Укажите способ (название графика), позволяющий определить качество обучения методом бинарной классификации

Ответ: ROC-кривая

Вопрос №8. Выберите способы борьбы с переобучением из следующего списка:
1. Ускорение вычислений с помощью графического процессора (GPU)
2. Регуляризация
3. Увеличение набора данных для обучения
4. Кросс-валидация

Вопрос №9. Назовите набор данных, необходимый для проверки качества обучения

Образцы заданий к практическим занятиям:

Задание 1. Работа с изображениями

- Откалибровать Web-камеру
- Записать себя с Web-камеры в файл с отрисованной с помощью cv2 рамкой вокруг найденного лица

Задание 2. Распознавание эмоций

- Возьмите за основу блокнот https://www.kaggle.com/ashishpatel26/tutorial-facial-expression-classification-keras
- Вместо 6 эмоций вернуться к 7
- Обучить сеть
- Превратить задачу классификации в задачу регрессии
- Улучить исходные результат за счет дополнения (augmentation): выравнивание по линии глаз, повороты, масштабирование и т.д.
- (Опционально) использовать Resnet (HINT: используйте не готовую Resnet, а найдите её реализацию, которая позволяет работать с изображениями произвольного размера. Если для выставить разрешение 96x96x3, то масштабирование от 48x48x1 будет произведено с минимальными потерями

Задание 3. Прогнозирование развития эпидемий

К данному моменту накоплено достаточное количество данных по распространению COVID-19:

https://github.com/CSSEGISandData/COVID-19\">CSSEGISandData/COVID-19

```
import pandas as pd
SITE = "https://raw.githubusercontent.com/"
REPO = "CSSEGISandData/COVID-19/master/
DIR = "csse_covid_19_data/csse_covid_19_time_series
FILE = "time_series_covid19_confirmed_global.csv"

df_cov = pd.read_csv(SITE + REPO + DIR + FILE)
rus_zero_offset = 10
rus_cov = df_cov[ df_cov['Country/Region'] == 'Russia'].drop(
  columns=['Province/State', 'Country/Region', 'Lat', 'Long']).stack()
  rus_cov.iloc[10:].head()
```

Также существуют хорошие модели развития эпидемий:
https://nplus1.ru/material/2019/12/26/epidemic-math
В наибольшей степени данную эпидемию (возможно!) описывают модели SEIR/SEIRS:
https://www.idmod.org/docs/hiv/model-seir.html
Ваша задача:
* отладить (найти и взять работающую) более-менее адекватную ситуации модель
* взять данные по нескольким странам из открытых данных
* обучить на этих данных (как на данных модели, так и данных) RNN-модель (лучше с хорошей памятью,например, стандартную LSTM)
* подавая её на вход данные, получить прогнозы динамики по другим странам

Основное, над чем надо подумать (и что будет оцениваться):
* Как отражать в RNN-модели разные подходы к течению эпидемии, которые используются в моделях SEIR/SEIRS.

В качестве дополнительного задания .... тем, кому интересно:
* Посмотрите на спектральные характеристики разных вариантов развития. Можете ограничиться преобразованием Фурье, можете использовать вейвлеты
https://pywavelets.readthedocs.io/en/latest/
* попробуйте использовать сверточные сети; иначе говоря, рассматривайте модели и данные эпидемий, как вход "изображения" для 1D-свертки и посмотрите, что у Вас получится

Задание 4. Работа с данными, представленными в виде графов

- Предложить свой dataset или выбрать из предложенных на https://mai.moscow/pages/viewpage.action?pageId=57802958
- Разработать объектную модель представления графа на java\scala\python
- Загрузить выбранный dataset
- Реализовать любой из алгоритмов bfs\dfs\Dijkstra

1.2. Промежуточная аттестация

Промежуточная аттестация осуществляется в форме зачет
В качестве средств, используемых на промежуточной аттестации предусматривается:
Билеты

1.3. Типовые задания для проведения промежуточной аттестации

Вопросы к зачету

1. Дайте определение большим данным, машинному обучению. Назовите области применения и примеры приложений.
2. Опишите схему разработки приложения для машинного обучения больших данных с примером.
3. Назовите основные этапы обработки сырых данных перед обучением. Приведите примеры.
4. Назовите основные классы алгоритмов машинного обучения с примерами.
5. Алгоритмы машинного обучения. Способы распараллеливания на примере «случайного леса».
6. Алгоритмы машинного обучения. Способы распараллеливания на примере градиентного бустинга.
7. Архитектура Apache Spark.
8. Глубокие нейронные сети. Пример
9. Анализ естественно-языковых текстов. Токенизация, стоп-слова, векторизация.
10. Анализ естественно-языковых текстов. TF-IDF.
11. Анализ больших данных, представленных в виде графа. Основные понятия, примеры.
12. Анализ больших данных, представленных в виде графа. Основные алгоритмы, примеры.
13. Назовите особенности анализа потоковых данных с примером.

2. КРИТЕРИИ ОЦЕНКИ ПО ДИСЦИПЛИНЕ

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	**2 (не зачтено)**	**3 (зачтено)**	**4 (зачтено)**	**5 (зачтено)**
Знания **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения **(виды оценочных средств: приведены в п. 1.2.)**	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки **(владения, опыт деятельности)** **(виды оценочных средств: приведены в п. 1.2..)**	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач