Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых» (ВлГУ)

Институт информационных технологий и радиоэлектроники

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
«Инжиниринг данных»

направление подготовки / специальность
09.04.01 «Информатика и вычислительная техника»

направленность (профиль) подготовки
Инженерия искусственного интеллекта

г. Владимир
2022

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью дисциплины «Инжиниринг данных» является формирование умений по применению научно-обоснованной комплексной методологии инжиниринга данных. Рассматриваются особенности работы с данными в различных форматах на языке Python. Уделяется внимание инструментам и технологиям загрузки данных из интернета и социальных сетей. Подробно изучаются методы очистки данных и соответствующие библиотеки на Python.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Инжиниринг данных» относится к обязательной части учебного плана.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции *(код, содержание индикатора*	Результаты обучения по дисциплине	
ОПК-3. Способен планировать и проводить комплексные исследования и изыскания для решения инженерных задач, относящихся к профессиональной деятельности, включая проведение измерений, планирование и постановку экспериментов, интерпретацию полученных результатов.	Знает: принципы, методы и средства анализа и структурирования профессиональной информации. Умеет: анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров. Имеет навыки: подготовки научных докладов, публикаций и аналитических обзоров с обоснованными выводами и рекомендациями.	ОПК-3. З-3. Сделать обзор основных методов статистической обработки и анализа результатов измерений. ОПК-3. У-2. Обоснованно выбрать необходимую аппаратуру и метод исследования для решения инженерных задач, относящихся к профессиональной деятельности ОПК-3. П-1. Выполнять в рамках поставленного задания экспериментальные комплексные научно-технические исследования и изыскания для решения инженерных задач в области профессиональной деятельности, включая обработку, интерпретацию и оформление 3 результатов ОПК-3. Д-1. Проявлять умение видеть детали, упорство, аналитические умения	Тестовые вопросы и задания, задания для самостоятельной работы, вопросы к зачету

4. ОБЪЕМ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 3 зачетные единицы, 108 часов

Тематический план
форма обучения – очная

№ п/п	Наименование тем и/или разделов/тем дисциплины	Семестр	Неделя семестра	Контактная работа обучающихся с педагогическим работником				Самостоятельная работа	Формы текущего контроля успеваемости, форма промежуточной аттестации *(по семестрам)*
				Лекции	Практические занятия	Лабораторные работы	*в форме практической подготовки*		
1	Работа с данными в Python	1	1-6	6	6		6	24	Рейтинг-контроль №1
2	Подготовка данных для систем машинного обучения.	1	7-12	6	6		6	24	Рейтинг-контроль №2
3	Параллельная и распределенная обработка данных.	1	13-18	6	6		6	24	Рейтинг-контроль №3
Всего за 1 семестр:				18	18			72	Зачет
Наличие в дисциплине КП/КР									
Итого по дисциплине				18	18			72	Зачет

Содержание лекционных занятий по дисциплине

1. **Работа с данными в Python**
   Библиотеки для работы с данными в различных форматах в Python: файлы CSV, JSON, HTML. Работа с базами данных в Python. Работа с изображениями, видео и звуковыми файлами. Форматы хранения больших данных и работа с ними: Parquet, Avro. Графы знаний.
2. **Подготовка данных для систем машинного обучения**
   Сбор данных и формирование набора данных для систем машинного обучения. Загрузка данных из интернет и социальных сетей. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных. Общедоступные платформы для хранения данных. Подход Data-Centric AI
3. **Параллельная и распределенная обработка данных**
   Архитектура центров обработки данных, кластеры для параллельных и распределенных вычислений. Экосистема для распределенного хранения и обработки больших объемов данных: Apache Hadoop, HDFS. Распределенная обработка данных в Apache Spark. Архитектура Apache Spark: Resilient Distributed Dataset (RDD), действия трансформации.

Работа с данными с использованием Spark DataFrame. Источники данных для Spark DataFrame. Обработка данных в Spark DataFrame. Использование SQL в Spark DataFrame.

Содержание практических занятий по дисциплине

1. Библиотеки для работы с данными в Python: numpy, pandas.
2. Работа с текстовыми файлами разных форматов в Python: CSV, JSON, HTML.
3. Работа с базами данных в Python.
4. Работа с изображениями, видео и звуковыми файлами в Python.
5. Работа с файлами для хранения больших данных в Python.
6. Работа с гр Создание собственных наборов данных в Python. Очистка и подготовка данныхафами знаний в Python.
7. Работа с данными в Apache Spark.
8. Использование SQL в Apache Spark.

4. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

5.1. Текущий контроль успеваемости

Рейтинг-контроль №1

Задание 1. Формат данных CSV может быть использован как замена: ● реляционным СУБД ● нереляционным СУБД ● документо-ориентированным СУБД

Задание 2. Что обычно выступает разделителем столбцов в формате CSV: ● Запятая ● Точка с запятой ● Перенос строки

Задание 3. Какие элементы не используются в формате JSON в качестве структурных элементов: ● Теги ● Кавычки ● Двоеточие

Рейтинг-контроль №2

Задание 5. Для какого языка программирования впервые использовался формат JSON:
● JavaScript
● Java
● Python 18

Задание 6. В каком виде хранятся данные в MongoDB:
● BSON
● JSON
● XML

Рейтинг-контроль №3

Задание 7. Выберите наиболее подходящие характеристики MongoDB:
● Ключ-значение, неструктурированность данных, нереляционные свойства
● Ключ-значение, неструктурированность данных, реляционные свойства
● SQL, неструктурированность данных, реляционные свойства

Задание 8. Преобразуйте следующий код в формат JSON:

```
<companies>
 <company>
 <company-id>7707040070</company-id>
 <name>Якорьбанк</name>
 <shortname>Якорьбанк</shortname>
```

<name-other>Якорьбанк, платёжное устройство</name-other>
<address>Россия, Республика Татарстан, Зеленодольский район, село Нурлаты, улица Гагарина, 46</address>
<phone>
<type>phone</type>
<number>+7 (800) 999-99-90</number>
</phone>
<url>http://www.yakorbank.ru/</url>
<working-time>будни 8:30-18:00, сб 9:00-14:30</working-time>
<rubric-id>184106974</rubric-id>
<actualization-date>23.09.2019</actualization-date>
<coordinates>
<lat>55.616051</lat>
<lon>48.295532</lon>
</coordinates>
</company>
</companies>

5.2. Промежуточная аттестация по итогам освоения дисциплины *(зачет)*

Вопросы к зачету:

1. Библиотека pandas в Python.
2. Работа с данными в формате CSV в Python.
3. Работа с данными в формате JSON в Python.
4. Работа с данными в формате HTML в Python.
5. Работа с изображениями в Python.
6. Работа с видео в Python.
7. Работа с аудио в Python.
8. Работа с Parquet в Python.
9. Работа с графами знаний в Python.
10. Этапы и инструменты создания наборов данных для машинного обучения.
11. Загрузка данных с Web-сайтов.
12. Загрузка данных из социальных сетей.
13. Методы и инструменты подготовки данных.
14. Методы и инструменты очистки данных.
15. Разметка данных.
16. Общедоступные платформы для хранения данных.
17. Архитектура центров обработки данных.
18. Кластеры для параллельных и распределенных вычислений.
19. Экосистема для распределенного хранения и обработки больших объемов данных: Apache Hadoop.
20. Распределенная файловая система HDFS.
21. Распределенная обработка данных в Apache Spark.
21
22. Работа с данными с использованием Apache Spark DataFrame.
23. Источники данных для Apache Spark DataFrame.
24. Обработка данных в Apache Spark DataFrame.
25. Использование SQL в Apache Spark DataFrame

5.3. Самостоятельная работа обучающегося

Подготовить собственный набор данных. Выберите задачу в одном из направлений

создания системы искусственного интеллекту (компьютерное зрение, обработка естественного языка) и подготовьте для этой задачи набор данных для обучения с учителем.

Соберите и очистите данные, проведите разметку. Готовый набор данных разместите на одной из общедоступных платформ для хранения данных по своему выбору. Подготовьте документацию к созданному набору данных.

Перечень задач, для которых рекомендуется подготовить набор данных:

- Классификация объектов на изображениях.
- Определение положения объектов на изображениях.
- Определение положения объектов в видео.
- Классификация текста на русском языке.
- Определение эмоциональной окраски текста на русском языке.

2. Создайте набор данных в Apache Spark и проведите его исследование с помощью Spark DataFrame API.

Схема данных выглядит следующим образом:

Онлайн-школа продает образовательные продукты: онлайн-курсы, книги, семинары и т.п.

Описание и стоимость продуктов содержится в таблице Products. Когда клиент что-то покупает, создается заказ, который заносится в таблицу Orders. Заказ может содержать несколько продуктов, перечень продуктов в заказах содержится в таблице Sales.

Таблица Products - продукты онлайн-школы:

- id - идентификатор продукта
- name - название продукта
- price - стоимость продукта

Таблица Oredrs - заказы:

- id - идентификатор заказа
- order_date - дата заказа
- customer_id - идентификатор заказчика (таблица с заказчиками не создается для упрощения примера)

Таблица Sales - продажи:

- product_id - идентификатор продукта, ссылка на таблицу Products, поле id
- order_date - идентификатор заказа, ссылка на таблицу Orders, поле id
- quantity - количество продуктов в заказе

Ноутбук в облачной платформе Colab с заготовкой кода для домашней работы – https://colab.research.google.com/drive/1MLilHIZ2CcBbCp_U7wKs_CAPmnx2F7Q6?usp=sharing

20

Задания для анализа:

- Выведите список продуктов, которые не были проданы ни разу
- Определите, сколько продуктов любого типа было продано по дням.
- Определить, какая выручка от продуктов любого типа была получена по дням.

Самостоятельная работа обучающихся заключается в самостоятельном изучении отдельных тем, практической реализации заданий самостоятельной работы по этим темам, выполнении контрольных работ. Контроль выполнения самостоятельной работы проводится при текущих контрольных мероприятиях и на промежуточной аттестации по итогам освоения дисциплины. Учебно-методическое обеспечение самостоятельной работы – основная

литература [1-3], дополнительная литература [1-2].

Фонд оценочных материалов (ФОМ) для проведения аттестации уровня сформированности компетенций обучающихся по дисциплине оформляется отдельным документом.

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Книгообеспеченность

Наименование литературы: автор, название, вид издания, издательство	Год издания	КНИГООБЕСПЕЧЕННОСТЬ
		Наличие в электронном каталоге ЭБС
Основная литература		
1. Садовникова, Н. А. Анализ временных рядов и прогнозирование / Садовникова Н. А. - Москва : Университет "Синергия", 2016. - 152 с. - ISBN 978-5-4257-0204-3. - Текст : электронный. Режим доступа : по подписке.	2016	https://www.studentlibrary.ru/book/ISBN9785425702043.html
2. Барский, А. Б. Введение в нейронные сети / Барский А. Б. - Москва : Национальный Открытый Университет "ИНТУИТ", 2016. - Текст : электронный. Режим доступа : по подписке.	2016	https://www.studentlibrary.ru/book/intuit_060.html
3. Хейдт, М. Изучаем pandas / Хейдт М. , пер. с анг. А. В. Груздева. - Москва : ДМК Пресс, 2018. - 438 с. - ISBN 978-5-97060-625-4. - Текст : электронный. Режим доступа : по подписке.	2018	https://www.studentlibrary.ru/book/ISBN9785970606254.html
Дополнительная литература		
1. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Флах П. - Москва : ДМК Пресс, 2015. - 400 с. - ISBN 978-5-97060-273-7. - Текст : электронный. Режим доступа : по подписке.	2015	https://www.studentlibrary.ru/book/ISBN9785970602737.html
2. Рашка, С. Python и машинное обучение : крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / Рашка С. - Москва : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0. - Текст : электронный. Режим доступа : по подписке.	2017	https://www.studentlibrary.ru/book/ISBN9785970604090.html

6.2. Периодические издания

1. Вестник компьютерных и информационных технологий ISSN 1810-7206.

2. Цифровая библиотека научно-технических изданий Института инженеров по электротехнике и радиоэлектронике (Institute of Electrical and Electronic Engineers (IEEE)) на английском языке – http://www.ieee.org/ieeexplore

6.3. Интернет-ресурсы

1. Academic Search Ultimate EBSCO publishing – http://search.ebscohost.com
2. eBook Collections Springer Nature – https://link.springer.com/
3. Гугл Академия – https://scholar.google.ru/

4. Электронно-библиотечная система «Лань» – https://e.lanbook.com/
5. Университетская библиотека ONLINE – https://biblioclub.ru/
6. Электронно-библиотечная система "Библиокомплектатор" (IPRbooks) http://www.bibliocomplectator.ru/available
7. Электронные информационные ресурсы Российской государственной библиотеки https://www.rsl.ru/
8. Научная электронная библиотека «КиберЛенинка» https://cyberleninka.ru/
9. Портал российского образования www.edu.ru
10. Портал российских электронных библиотек www.elbib.ru
11. Научная электронная библиотека www.eLibrary.ru
12. Научная библиотека ВлГУ library.vlsu.ru
13. Учебный сайт кафедры ИСПИ ВлГУ https://ispi.cdo.vlsu.ru
14. Электронная библиотечная система ВлГУ https://vlsu.bibliotech.ru/
15. М.В. Ронкин. Курс Time Series Analysis. URL: https://github.com/MVRonkin/Time-Series-Analysis-Lectures-and-Workshops
16. Примеры использования библиотеки SKTimes. URL: https://github.com/sktime/sktime-tutorial-pydata-amsterdam-2020
17. Практический Анализ временных рядов. URL: https://github.com/nmmarcelnv/PracticalTimeSeries
18. Список открытых ресурсов по анализу временных рядов с использованием методов глубокого обучения нейронных сетей. URL: https://github.com/Alro10/deep-learning-time-series
19. Список открытых ресурсов по анализу временных рядов. URL: https://github.com/bifeng/Awesome-time-series
20. Список библиотек анализа временных рядов для языка программирования Python. URL: https://github.com/MaxBenChrist/awesome_time_series_in_python
21. Ресурс, посвященный методам и наборам данных для классификации временных рядов. URL: http://timeseriesclassification.com/index.php
22. Репозиторий, связанный с книгой Practical Time Series Analysis. URL: https://github.com/PracticalTimeSeriesAnalysis/BookRepo
23. Архив наборов данных для анализа временных рядов. URL: https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Для реализации данной дисциплины имеются специальные помещения для проведения занятий: занятий лекционного типа, практических занятий, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы. Практические занятия проводятся в компьютерном классе, оборудованном мультимедийным проектором с экраном и обеспеченным доступом в Интернет.

Перечень используемого лицензионного программного обеспечения:

- Операционная система Microsoft Windows 10
- Офисный пакет Microsoft Office 2016
- Бесплатное программное обеспечение (Python – https://www.python.org/, PyTorch - https://pytorch.org/, TensorFlow, Keras - https://www.tensorflow.org/, Sktime - https://www.sktime.org/en/v0.4.2/, Pandas - https://pandas.pydata.org/, Anaconda solution - https://www.anaconda.com/, Веб - среда разработки для языка программирования Python: google colab - https://colab.research.google.com/)

Рабочую программу составил Куликов К.В. зав. каф. ВТиСУ
(ФИО, должность, подпись)

Рецензент
(представитель работодателя) _____ Генеральный директор ООО "Диаграмма" Протягов И.В.

Программа рассмотрена и одобрена на заседании кафедры ВТ и СУ
Протокол № 1 от 29 августа 2022 года
Заведующий кафедрой Куликов К.В. _____

Рабочая программа рассмотрена и одобрена
на заседании учебно-методической комиссии направления 09.04.01 информатика и вычислительная техника
Протокол № 1 от 29 августа 2022 года
Председатель комиссии Куликов К.В. зав. каф. ВТиСУ_____

ЛИСТ ПЕРЕУТВЕРЖДЕНИЯ
РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

Рабочая программа одобрена на 20____ / 20____ учебный года
Протокол заседания кафедры № ______ от _________ года
Заведующий кафедрой____

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

в рабочую программу дисциплины

Инжиниринг данных

образовательной программы направления подготовки *09.04.01«Информатика и вычислительная техника»*, направленность: *Инженерия искусственного интеллекта (магистратура*)

Номер изменения	Внесены изменения в части/разделы рабочей программы	Исполнитель ФИО	Основание (номер и дата протокола заседания кафедры)

Заведующий кафедрой _____/ _____