**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  29.12.2022 г. |

**Макрушин С.В., Блохин Н.В.**

**ТЕХНОЛОГИИ ОБРАБОТКИ ДАННЫХ**

**Рабочая программа дисциплины**

для студентов, обучающихся по направлению подготовки

09.03.03 - Прикладная информатика,

ОП «Инженерия данных»,

ОП «Прикладные информационные системы в экономике и финансах»,

*Рекомендовано Ученым советом*

*Факультета информационных технологий и анализа больших данных*

*(протокол №27 от 15.12.2022г.)*

*Одобрено Советом учебно-научного*

*Департамента анализа данных и машинного обучения*

*(протокол №6 от 13.12.2022 г.)*

**Москва 2022**

**Оглавление**

[1. Наименование дисциплины 2](#_Toc74604615)

[2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине 2](#_Toc74604616)

[3. Место дисциплины в структуре образовательной программы 3](#_Toc74604617)

[4. Объем дисциплины(модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся 3](#_Toc74604618)

[5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий 4](#_Toc74604619)

[5.1. Содержание дисциплины 4](#_Toc74604620)

[5.2. Учебно-тематический план 7](#_Toc74604621)

[5.3. Содержание семинаров, практических занятий 9](#_Toc74604622)

[6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине 10](#_Toc74604623)

[6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы 10](#_Toc74604624)

[6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю 12](#_Toc74604625)

[7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине 13](#_Toc74604626)

[8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины 15](#_Toc74604627)

[9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины 16](#_Toc74604628)

[10. Методические указания для обучающихся по освоению дисциплины 17](#_Toc74604629)

[11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем 18](#_Toc74604630)

[12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине. 18](#_Toc74604631)

# 1. Наименование дисциплины

«Технологии обработки 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-6 | Способность организовывать поиск и сбор информации, ее хранение в структурированном виде, проектировать и реализовывать реляционные и нереляционные базы и хранилища данных | Демонстрирует знание основ реляционных баз данных, нормализации данных, ACID, CRUD, ORM, использует транзакции. | **Знать:** способы хранения данных с отношениями «один ко многим» и «один к одному».  **Уметь:** с помощью языка Python сохранять и читать данные с отношениями «один ко многим» и «один к одному». |
| Демонстрирует знание различных технологий хранения данных: реляционные и нереляционные базы данных, документарные хранилища, извлекает данные из разных источников и в разных форматах, в том числе программно. | **Знать:** различные технологии сериализации, хранения и чтения данных.  **Уметь:** с помощью языка Python сохранять и читать данные в популярных универсальных форматах. |
| Проектирует хранилища данных исходя из их назначения и характера данных, выбирает инструментальное и архитектурное решение, физическую и логическую схему данных и обосновывает свой выбор. | **Знать:** подходы к организации структуры данных в форматах CSV, XML и JSON.  **Уметь:** с помощью языка Python сохранять и читать данные в форматах CSV, XML и JSON. |

# 3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки данных» относится к Общепрофессиональному циклу дисциплин по направлению подготовки 09.03.03 - Прикладная информатика, ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах».

Изучение дисциплины «Технологии обработки данных» основывается на сумме знаний, полученных при изучении дисциплины «Алгоритмы и структуры данных в языке Python». Для изучения данной дисциплины студент должен обладать базовыми знаниями в области информационных технологий и программирования, навыками программирования на языке Python.

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

*ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах»*

*Очная форма обучения / очно-заочная форма обучения*

|  |  |  |
| --- | --- | --- |
| Вид учебной работы по  дисциплине | Всего  (в з.е. и часах) | Семестр 2/2  (в часах) |
| **Общая трудоёмкость**  **дисциплины** | 4/144 | 144 |
| ***Контактная работа-***  ***Аудиторные занятия*** | ***50/18*** | ***50/18*** |
| Лекции | 16/0 | 16/0 |
| Семинары, практические занятия | 34/18 | 34/18 |
| ***Самостоятельная работа*** | ***94/126*** | ***94/126*** |
| Вид текущего контроля | Контрольная  работа | Контрольная  работа |
| Вид промежуточной аттестации | Зачет | Зачет |

*ОП «Прикладные информационные системы в экономике и финансах»*

*Заочная форма обучения (Институт онлайн-образования)*

|  |  |  |
| --- | --- | --- |
| Вид учебной работы по  дисциплине | Всего  (в з.е. и часах) | Семестр 2  (в часах) |
| **Общая трудоёмкость**  **дисциплины** | 4/144 | 144 |
| ***Контактная работа-***  ***Аудиторные занятия*** | ***16*** | ***16*** |
| Лекции | 4 | 4 |
| Семинары, практические занятия | 12 | 12 |
| ***Самостоятельная работа*** | ***128*** | ***128*** |
| Вид текущего контроля | контрольная  работа | контрольная  работа |
| Вид промежуточной аттестации | Зачет | Зачет |

# 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. Библиотека NumPy и Pandas.**

В рамках темы рассматривается технологический стек Python для обработки и анализа данных, возможности Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными. Универсальные функции и применение функций по осям в NumPy. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры Маскирование и прихотливое индексирование в NumPy.

В рамках темы рассматриваются возможности библиотеки Pandas. Организация Pandas DataFrame и организация индексации для DataFrame и Series; применение универсальных функций и работа с пустыми значениями в Pandas. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры. Рассматривается операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение».

**Тема 2. Использование различных форматов файлов в задачах обработки данных.**

В рамках темы рассматриваются принципы работы с файлами, файлы и операционные системы. Специфика текстовых и бинарных файлов.

В рамках темы рассматривается задача сериализации и десериализации данных и использование различных форматов файлов для ее решения. Описание формата файла JSON и пример описания данных в этом формате и взаимодействия с ним в Python.

В рамках темы рассматриваются формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM, работа с ними с помощью библиотеки BeautifulSoup.

В рамках темы рассматривается проблематика форматов файлов для хранения и обработки больших данных. Форматы файлов NPY и HDF: общая характеристика, пример взаимодействия с данными этих форматов в Python.

**Тема 3. Взаимодействие c табличными данными в приложениях обработки данных.**

В рамках темы рассматривается формат файлов CSV, представление данных в этом формате и взаимодействие с ним в Python.

В рамках темы рассматриваются возможности использования Excel для внешних приложений обработки данных. Взаимодействие с Excel из Python с помощью библиотеки XLWings: принципы работы и примеры использования.

**Тема 4. Визуализация данных.**

В рамках темы рассматриваются основы работы с библиотекой matplotlib: организация системы координат, оформление осей, цвета и цветовые карты в matplotlib, стили линий и маркеры. Pyplot и объектно-ориентированный интерфейс matplotlib. Управление фигурами и создание множества графиков на одном рисунке. Различные типы графиков.

В рамках темы рассматривается визуализация данных с помощью библиотеки Pandas: набор методов для построения графиков, реализованный в структурах Series и DataFrame.

В рамках темы проводится введение в разведочный анализ данных: типы признаков, анализ распределений, анализ мер центральной тенденции и поиск выбросов, анализ взаимного распределения и парных корреляций. Проведение разведочного анализа данных с помощью библиотеки Seaborn.

**Тема 5. Работа со строками в приложениях обработки данных.**

В рамках темы рассматриваются возможности python по форматированию строк: %-форматирование, метод format, f-строки.

В рамках темы рассматриваются основы работы с регулярными выражениями: базовый синтаксис, примеры. Модуль ***re*** в Python. Примеры использования регулярных выражений.

В рамках темы рассматривается использования хэширования при работе со строками. Строки в библиотеке numpy.

**Тема 6. Введение в обработку текста на естественном языке в задачах обработки данных.**

В рамках темы рассматриваются сегментация и токенизация текста на естественном языке, стеммминг и лемматизация, примеры на Python. Использование мемоизации на примере работы со строками. Расстояние Левеншнтейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python. Векторное представление текста на естественном языке: общий алгоритм подходов TF; TF-IDF.

**Тема 7. Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba.**

В рамках темы рассматривается профилирование реализации алгоритмов на Python, принципы решения задачи оптимизации производительности алгоритма. Библиотека Numba: принципы работы, базовые примеры использования. Векторизация в numpy: ключевые параметры функции, примеры применения, использование обобщенной сигнатуры функции.

## 5.2. Учебно-тематический план

*ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах»*

Очная форма обучения / очно-заочная форма обучения

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| №  п/п | **Наименование темы (раздела) дисциплины** | **Трудоемкость в часах** | | | | | **Формы текущего контроля успеваемости** |
| **Всего** | **Контактная работа-**  **Аудиторная работа** | | | **Самостоятельная работа** |
| Общая, в т.ч.: | Лекции | Семинары, практические занятия |
| 1 | Библиотека NumPy и Pandas | 28/22 | 14/4 | 4/0 | 10/4 | 14/18 | Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.  Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям. |
| 2 | Использование различных форматов файлов в задачах обработки данных. | 20/22 | 6/4 | 2/0 | 4/4 | 14/18 |
| 3 | Взаимодействие c табличными данными в приложениях обработки данных. | 20/20 | 6/2 | 2/0 | 4/2 | 14/18 |
| 4 | Визуализация данных | 20/20 | 6/2 | 2/0 | 4/2 | 14/18 |
| 5 | Работа со строками в приложениях обработки данных | 20/20 | 6/2 | 2/0 | 4/2 | 14/18 |
| 6 | Введение в обработку текста на естественном языке в задачах обработки данных | 20/20 | 6/2 | 2/0 | 4/2 | 14/18 |
| 7 | Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba. | 16/20 | 6/2 | 2/0 | 4/2 | 10/18 |
|  | В целом по дисциплине | 144 | 50/18 | 16/0 | 34/18 | 94/126 | Согласно учебному плану: контрольная работа |
|  | Итого в % |  | 35/13 | 32/- | 68/100 | 65/87 |  |

*ОП «Прикладные информационные системы в экономике и финансах»*

Заочная форма обучения (Институт онлайн-образования)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| №  п/п | **Наименование темы (раздела) дисциплины** | **Трудоемкость в часах** | | | | | **Формы текущего контроля успеваемости** |
| **Всего** | **Контактная работа-**  **Аудиторная работа** | | | **Самостоятельная работа** |
| Общая, в т.ч.: | Лекции | Семинары, практические занятия |
| 1 | Библиотека NumPy и Pandas | 24 | 6 | 2 | 4 | 18 | Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.  Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям. |
| 2 | Использование различных форматов файлов в задачах обработки данных. | 23 | 5 | 1 | 4 | 18 |
| 3 | Взаимодействие c табличными данными в приложениях обработки данных. | 23 | 5 | 1 | 4 | 18 |
| 4 | Визуализация данных | 18 | 0 | 0 | 0 | 18 |
| 5 | Работа со строками в приложениях обработки данных | 18 | 0 | 0 | 0 | 18 |
| 6 | Введение в обработку текста на естественном языке в задачах обработки данных | 18 | 0 | 0 | 0 | 18 |
| 7 | Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba. | 20 | 0 | 0 | 0 | 20 |
|  | В целом по дисциплине | 144 | 16 | 4 | 12 | 128 | Согласно учебному плану: контрольная работа |
|  | Итого в % |  | 11 | 2 | 12 | 89 |  |

## 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** |
| Библиотека NumPy и Pandas | * Технологический стек Python для обработки и анализа данных * Возможности Python как glue language * Организация массивов в NumPy: хранение данных, создание массивов * Принципы реализации операций с едиными исходными данными. Универсальные функции и применение функций по осям в NumPy. * Организация Pandas DataFrame и организация индексации для DataFrame и Series. * Применение универсальных функций и работа с пустыми значениями в Pandas. * Объединение данных из нескольких Pandas DataFrame: общая логика и примеры.   8[1], 9[9], 9[10] | Интерактивная форма, работа на компьютере |
| Использование различных форматов файлов в задачах обработки данных | * Формат файлов Pickle, представление данных в этом формате и взаимодействие с ним в Python. * Формат файлов JSON, представление данных в этом формате и взаимодействие с ним в Python. * Формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM * Работа с XML с помощью библиотеки BeautifulSoup.   8[1], 8[2], 9[3], 9[4] | Интерактивная форма, работа на компьютере |
| Взаимодействие c табличными данными в приложениях обработки данных. | * Взаимодействие с Excel из Python с помощью библиотеки XLWings. * Формат файлов CSV, представление данных в этом формате и взаимодействие с ним в Python   8[1], 8[2] | Интерактивная форма, работа на компьютере |
| Визуализация данных | * Построение визуализаций с помощью библиотеки matplotlib * Построение визуализаций с помощью библиотеки pandas * Построение визуализаций с помощью библиотеки seaborn   8[1], 9[13], 9[15], 9[16] | Интерактивная форма, работа на компьютере |
| Работа со строками в приложениях обработки данных | * Основы работы с регулярными выражениями: базовый синтаксис, примеры. * Модуль re в Python.   8[1], 8[2], 9[4] | Интерактивная форма, работа на компьютере |
| Введение в обработку текста на естественном языке в задачах обработки данных. | * Сегментация и токенезация текста на естественном языке, стеммминг и лемматизация, примеры на Python. * Расстояние Левеншнтейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python.   8[1], 8[2], 9[4], 9[5], 9[6] | Интерактивная форма, работа на компьютере |
| Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba | * профилирование реализации алгоритмов на Python * принципы решения задачи оптимизации производительности алгоритма * Библиотека Numba: принципы работы, базовые примеры использования.   8[1], 8[2], 9[1], 9[2], 9[3] | 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** |
| Библиотека NumPy и Pandas | * Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры. * Маскирование и прихотливое индексирование в NumPy. * Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение». | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Использование различных форматов файлов в задачах обработки данных | * Формат файлов NPY, представление данных в этом формате и взаимодействие с ним в Python. * Формат файлов HDF, представление данных в этом формате и взаимодействие с ним в Python. | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Взаимодействие c табличными данными в приложениях обработки данных. | * Продвинутые операции с Excel из Python с помощью библиотеки XLWings. | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Визуализация данных | * Построение трехмерных графиков   Продвинутая работа с цветовыми картами | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Работа со строками в приложениях обработки данных | * Использования хэширования при работе со строками. * Строки в библиотеке numpy. | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Введение в обработку текста на естественном языке в задачах обработки данных. | * Использование мемоизации на примере работы со строками. * Векторное представление текста на естественном языке: общий алгоритм подходов TF; TF-IDF. | Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач. |
| Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba | * Векторизация в numpy: ключевые параметры функции, примеры применения * Использование обобщенной сигнатуры функции в numpy и numba. | 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 контрольной работе***

1. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными
2. Универсальные функции и применение функций по осям в NumPy
3. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры
4. Маскирование и прихотливое индексирование в NumPy
5. Векторизация в numpy: ключевые параметры функции, примеры применения, использование обобщенной сигнатуры функции

***Примерные задания контрольной работы***

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

*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-6**  Способность организовывать поиск и сбор информации, ее хранение в структурированном виде, проектировать и реализовывать реляционные и нереляционные базы и хранилища данных | Демонстрирует знание основ реляционных баз данных, нормализации данных, ACID, CRUD, ORM, использует транзакции. | **Знать:** способы хранения данных с отношениями «один ко многим» и «один к одному».  **Уметь:** с помощью языка Python сохранять и читать данные с отношениями «один ко многим» и «один к одному». | Сохранить информацию телефонной книги, содержащей отношения «один к одному» и «один ко многим» в формате XML. |
| Демонстрирует знание различных технологий хранения данных: реляционные и нереляционные базы данных, документарные хранилища, извлекает данные из разных источников и в разных форматах, в том числе программно. | **Знать:** различные технологии сериализации, хранения и чтения данных.  **Уметь:** с помощью языка Python сохранять и читать данные в популярных универсальных форматах. | Сериализовать информацию телефонной книги в формате Pickle и XML. |
| Проектирует хранилища данных исходя из их назначения и характера данных, выбирает инструментальное и архитектурное решение, физическую и логическую схему данных и обосновывает свой выбор. | **Знать:** подходы к организации структуры данных в форматах CSV, XML и JSON.  **Уметь:** с помощью языка Python сохранять и читать данные в форматах CSV, XML и JSON. | Сохранить и прочитать информацию телефонной книги используя формат JSON. |

***Примерные вопросы для подготовки к зачету***

1. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными
2. Универсальные функции и применение функций по осям в NumPy
3. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры
4. Маскирование и прихотливое индексирование в NumPy
5. Векторизация в numpy: ключевые параметры функции, примеры применения, использование обобщенной сигнатуры функции
6. Numba: принципы работы, базовые примеры использования
7. Организация Pandas DataFrame и организация индексации для DataFrame и Series
8. Применение универсальных функций и работа с пустыми значениями в Pandas
9. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры
10. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение»
11. Специфика текстовых и бинарных файлов, форматы файлов CSV и Pickle, представление данных в этих форматах и взаимодействие с ними в Python
12. Задача сериализации и десериализации, описание формата файла JSON и пример описания данных в этом формате и взаимодействия с ним в Python
13. Формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM, работа с ними с помощью библиотеки BeautifulSoup
14. Форматы файлов NPY и HDF общая характеристика, пример взаимодействие с данными этих форматов в Python
15. Взаимодействие из Python с базой данных на примере API SQLite, базовые возможности работы с транзакциями
16. Взаимодействие с Excel из Python с помощью XLWings: принципы работы и примеры использования
17. Основы работы с регулярными выражениями: базовый синтаксис, примеры использования модуля re в Python
18. Сегментация и токенезация текста на естественном языке, стеммминг и лемматизация, примеры на Python
19. Расстояние Левеншнтейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python
20. Векторное представление текста на естественном языке: общий алгоритм подходов TF; TF-IDF

# 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: 17.01.2023). – Текст: электронный.

***Дополнительная литература:***

1. 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: 17.01.2023). – 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. Pylru 1.0.9 [Электронный ресурс]: сайт. – Режим доступа: <https://pypi.python.org/pypi/pylru>

11. Python Data Analysis Library [Электронный ресурс]: сайт. – Режим доступа: <http://pandas.pydata.org/>

12. Python Documentation [Электронный ресурс]: сайт. – Режим доступа: <http://python.org/doc/>

1. Python Standard Library [Электронный ресурс]: сайт. – Режим доступа: <https://docs.python.org/2/library/>
2. Scikit-learn Machine Learning in Python [Электронный ресурс]: сайт. – Режим доступа: <http://scikit-learn.org>
3. Официальный сайт продукта <https://www.python.org/>
4. Каталог курсов Интернет Университета Информационных Технологий <http://www.intuit.ru/>
5. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
6. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
7. Pandas User Guide http://pandas.pydata.org/pandas-docs/stable/

**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.