Программа учебной дисциплины
«Анализ данных в Python»

Утверждена
Академическим советом ОП
Протокол № 2.6.4-02/01 от 29. 08. 2019 г.

Разработчик	Рогович Татьяна Владимировна, Приглашенный преподаватель, Департамент больших данных и информационного поиска
Число кредитов	4
Контактная работа (час.)	62
Самостоятельная работа (час.)	90
Курс, Образовательная программа	4 (Б) курс, Политология
Формат изучения дисциплины	С использованием онлайн курса: https://www.datacamp.com/

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. Развитие и закрепление навыков программирования на языке Python.
2. Формирование и развитие навыков работы со специализированными библиотеками для обработки, визуализации и анализа данных (pandas, numpy, scipy, sklearn, plotly, matplotlib).
3. Развитие навыков работы с данными: сбор, обработка, визуализация, разведывательный анализ.
4. Освоение терминологии области машинного обучения и знакомство с базовыми алгоритмами
5. Развитие навыков постановки исследовательской задачи и тестирования гипотез с помощью количественных методов.
6. Развитие навыков презентации полученных результатов (оформление отчета о проделенной работе и устная защита исследования).

Планируемые результаты обучения (ПРО):

1. Уверенно пользоваться языком Python для решения аналитических задач
2. Загружать данные в pandas и работать с ними (фильтрация, агрегация, заполнение пропущенных значений)
3. Умение подсчитывать описательные статистики, оценивать распределения, интерпретировать корелляции

4. Умение выбирать корректные графики для визуализации данных, уметь кастомизировать их внешний вид, интерпретировать графики
5. Создавать интерактивные визуализации с помощью plotly
6. Собирать данные через API. Преобразовывать формат json в таблицу
7. Собирать данные с помощью web-scraping, парсить данные и сохранять их в табличном виде
8. Определять тип задачи машинного обучения, выбирать корректные модели для ее решения, осуществлять подбор параметров и выбирать лучшую модель
9. Проводить разведывательный анализ данных
10. Решать задачи машинного обучения от постановки исследовательского вопроса до интерпретации результатов
11. Решать простые задачи классификации, регрессии и кластеризации
12. Работать с сайтом соревнований по машинному обучению kaggle
13. Собирать и подготавливать данные для текстового анализа. Проводить стандартизацию текста. Решать задачи классификации и кластеризации для текстовых данных
14. Подготавливать данные для сетевого анализа и строить социальные графы
15. Иметь общее представление о принципах работы нейроных сетей. Уметь использовать готовые нейронные сети для анализа собственных данных

Пререквизиты:

1. Успешное прохождение курса “Основы программирования в Python” или его аналога или самостоятельное освоение тем, заявленных в его программе. Во втором случае будет проведен тест или собеседование с преподавателем.

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах**	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
	лк		
	см		
	onl/cp		
Введение в анализ данных: Pyhton для анализа данных, алгоритмы, введение в библиотеки pandas и numpy.	8	№: 1, 2, 3, 9, 12.	КР1, КР2, ДБ, Проект.
	10		
	20		
Визуализация данных	4	№: 4, 5.	ДЗ1, ДБ, Проект.
	4		
	10		
Сбор данных	4	№: 6, 7.	ДЗ1, ДБ, Проект.
	2		
	8		

Машинное обучение	8 8 30	№: 8, 9, 10, 11, 12.	ДБ, Проект.
Текстовый анализ	4 4 10	№: 13.	ДЗ2, ДБ.
Сетевой анализ	2 2 8	№: 14.	ДБ.
Нейронные сети	2 0 4	№: 15.	ДБ.
Часов по видам учебных занятий:	32 30 90		
Итого часов:	152		

Содержание разделов дисциплины:

1. **Введение в анализ данных: Pyhton для анализа данных, алгоритмы, введение в библиотеки pandas и numpy.**
   Преимущества использования Python для анализа данных по сравнению с другими инструментами. Прикладные задачи политологии, для решения которых подходит инструментарий Python. Обзор библиотек и инструментов. Программирование на Python: вспоминаем типы данных, основные структуры, методы и функции, условные операторы, циклы, списковые включения, функцию map() и анонимные функции, отладку кода (try/except). Алгоритмы: оптимизация и сложность на примере алгоритмов сортировок. Библиотека numpy: векторы и массивы, специальные типы данных. Библиотека pandas. Основы работы с датафреймами: загрузка, очистка, фильтрация, группировка и агрегация. Описательные статистики, распределения, разведывательный анализ, работа с пропущенными значениями, постановка гипотезы. Работа с Kaggle. Дополнительные материалы: Разделы 1-4 курса "Python для извлечения и обработки данных" https://openedu.ru/course/hse/PYTHON/
2. **Визуализация данных**
   Виды графиков, их корректное использование и интерпретация. Принципы хорошей визуализации, основные ошибки при вузализации данных. Основы визуализации в matplotlib. Оформление и кастомизация графиков. Визуализации отфильтрованных и сгруппированных данных. Создание интерактивных визуализаций в Plotly. Дополнительные материалы: Раздел 9 курса "Python для извлечения и обработки данных" https://openedu.ru/course/hse/PYTHON/

3. **Сбор данных**
   Сбор данных из открытых источников: web-scraping, работа с API, парсинг текста и таблиц, сохранение файлов. Дополнительные материалы: Разделы 6-8 курса "Python для извлечения и обработки данных" https://openedu.ru/course/hse/PYTHON/
4. **Машинное обучение**
   Введение в машинное обучение: терминология, постановка исследовательского вопроса и проверка гипотезы. Виды задач машинного обучения. Baseline модели. Numpy: операции с векторами и матрицами. Задачи регрессии. Линейная регрессия. Задачи классификации. Реализация алгоритма kNN. Логистическая регрессия. Решающие деревья. Случайный лес. Работа с Kaggle. Обучение без учителя. Кластеризация. Решаем Kaggle кейс: от гипотезы до submission.
5. **Текстовый анализ**
   Сбор, обработка и очистка текста для анализа. Текстовый анализ: классификация, семантический анализ. Байесовские модели. Выявление топиков (LDA). Дополнительные материалы: Раздел 5 курса "Python для извлечения и обработки данных" https://openedu.ru/course/hse/PYTHON/ или Natural Language Processing Fundamentals in Python https://www.datacamp.com/courses/natural-language-processing-fundamentals-in-python
6. **Сетевой анализ**
   Сетевой анализ: кейсы, описательные статистики. Сбор и подготовка данных. Построение социального графа. Дополнительный материал: Network Analysis in Python https://www.datacamp.com/courses/network-analysis-in-python-part-1 https://www.datacamp.com/courses/network-analysis-in-python-part-2
7. **Нейронные сети**
   Нейронные сети. Области применения. Использование существующих решений для собственных задач.

3. Оценивание

- **КР1**, Не блокирующее, Контрольная работа
  Контрольная работа во время семинара. (раздел 1). Проверяется навык программирования на Python.
- **КР2**, Не блокирующее, Контрольная работа
  Контрольная работа во время семинара (раздел 1). Проверяется навык работы в библиотеке pandas (агрегация, фильтрация данных, создание новых переменных, работа с описательными статистиками, распределениями).
- **ДЗ1**, Не блокирующее, Домашнее задание
  Домашнее задание. Проверяется умение собирать данные из интернета, создавать новые переменные, объединять данные из разных источников, строить и интерпретировать интерактивные визуализации, описывать выводы).

- **ДЗ2**, Не блокирующее, Домашнее задание
  Домашнее задание. Проверяется умение создавать собственный датасет для текстового анализа (сбор данных, присвоение лейблов), умение проводить обработку и стандартизацию текста, применять модели для текстового анализа и интерпретировать результаты.
- **ДБ**, Не блокирующее, Домашнее задание
  Дополнительные баллы. У студентов есть возможность получить до трех дополнительных баллов за выполнение необязательных заданий в течение семестра. Дополнительные баллы учитываются в итоговой оценке до округления с весом 1. Обратите внимания, что дополнительные задания, выполненные в рамках онлайн курса засчитываются только при условии прохождения исключительно с корпоративного почтового адреса студента. Подключение студентов к онлайн курсу на платформе НПОО (https://openedu.ru/) производит Дирекция по онлайн обучению НИУ ВШЭ по заявке администратора учебного офиса образовательной программы. Скрытая сессия для студентов ВШЭ автоматически появляется в личном аккаунте на платформе. Регистрироваться на открытую сессию для всех желающих слушателей нельзя. На платформе DataCamp слушатели приглашаются в специальную сессию, созданную преподавателем курса.
- **Проект**, Не блокирующее, Экзамен (устный)
  Самостоятельно выполненный проект по машинному обучению на данных по выбору и его защита.

Формула округления: Стандартное арифметическое округление

Вид формулы оценивания: Линейная
Формула оценивания:

Окончательная оценка = Округление(0.7 * ((КР1 + КР2 + ДЗ1 + ДЗ2) / 4) + 0.3 * Проект + ДБ)

Преподаватель оставляет за собой право устроить устную защиту любой из форм контроля.

Если во время проверки контрольной работы, проектного задания или домашнего задания выявлен факт нарушения академической этики, то студент получает оценку «0» за данную работу. Работа студента, предоставившего работу для списывания, тоже аннулируется. Преподаватель оставляет за собой право применить дисциплинарное взыскание к обоим студентам.

Если во время написания контрольных работ студент нарушает правила проведения контрольной работы, то студент получает за эту работу оценку «0». К нарушению правил относятся: списывание, коммуникация с другими студентами, использование телефона/умных часов/планшета и т.д., использование материалов, не оговоренных преподавателем.

Домашнее задание должно быть сдано до установленного дедлайна. В случае сдачи в течение суток после дедлайна, оценка снижается на 1 балл. В случае сдачи в течение недели после дедлайна, оценка снижается на 2 балла. Работы, сданные позже, не принимаются и за них выставляется оценка «0».

Проектная работа сдается к дедлайну перед экзаменом (устная защита). Проектные работы, сданные после дедлайна, не принимаются. Студенту ставится неявка на экзамен при отсутствии сданной вовремя проектной работы или при неявке на устную защиту (даже при наличии сданной работы).

Если студент отсутствовал во время контрольной работы или экзамена или не смог вовремя сдать домашнее задание по причине болезни или по другим уважительным причинам, подтвержденных учебным офисом, он получает право на пересчет итоговой оценки без учета элемента контроля, пересдачу экзамена или продление дедлайна. В других случаях отсутствие студента считается неявкой и он получает оценку «0» за контрольную работу.

4. Примеры оценочных средств

КР1: задачи по программированию. Например, как в этом курсе: https://www.coursera.org/learn/python-osnovy-programmirovaniya

КР2: будет дан датасет, для которого нужно будет выполнить ряд операций (фильтрация, агрегация, создание новых переменных, поиск наблюдений, подсчет описательных статистик, сделать визуализации и т.д.).

Итоговый проект: проект по данным на ваш выбор (оформленный ноутбук с описанием данных, разведывательным анализом, постановкой гипотезы, ее тестированием и выводами).

Пример проекта: https://drive.google.com/file/d/1wiHukFIUzjcz9X49vQRLXgkUi7ga5C7e/view?u...

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	Уэс, М. Python и анализ данных / М. Уэс ; перевод с английского А.А. Слинкин. — Москва : ДМК Пресс, 2015. — 482 с. — ISBN 978-5-97060-315-4. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/73074
2	Plotly Python Open Source Graphing
3	Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани ; перевод с английского С.Э. Мастицкого. —

	Москва : ДМК Пресс, 2017. — 456 с. — ISBN 978-5-97060-495-3. — URL: https://e.lanbook.com/book/93580
4	*М. Бонцанини* Анализ социальных медиа на Python. Извлекайте и анализируйте данные из всех уголков социальной паутины на Python, перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. — 288 с. — ISBN 978-5-97060-574-5. — Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/108129

5.2. Рекомендуемая дополнительная литература
Не требуется

5.3. Программное обеспечение

п/п	**Наименование**	**Условия доступа/скачивания**
1	Microsoft Windows 7 Professional RUS Microsoft Windows 8.1 Professional RUS Microsoft Windows 10	*Из внутренней сети университета (договор)*
2	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3	интерпретатор языка Python 3.7 (или более новых версий)	*Свободное лицензионное соглашение*
4	Anaconda, бесплатный дистрибутив языков программирования Python и R для научных вычислений	*Свободное лицензионное соглашение*

5.4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

п/п	**Наименование**	**Условия доступа/скачивания**
	Профессиональные базы данных, информационно-справочные системы	
1	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
2	Открытое образование https://openedu.ru/	Подключение студентов к онлайн курсу на платформе НПОО (https://openedu.ru/) производит Дирекция по онлайн обучению НИУ ВШЭ по заявке администратора учебного офиса образовательной программы.
3	DataCamp https://www.datacamp.com/	Подключение через сессию созданную преподавателем по бесплатной академической лицензии
	Интернет-ресурсы (электронные образовательные ресурсы)	
1	Открытое образование	URL: https://openedu.ru/

5.5. Материально-техническое обеспечение дисциплины
Учебные аудитории для лекционных по дисциплине обеспечивают использование

и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:
- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.
Учебные аудитории для семинарских и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

Компьютерные классы оборудованы ПЭВМ с доступом в Интернет, операционными системами и программным обеспечением, необходимыми для освоения дисциплины. При необходимости допускается замена оборудования его виртуальными аналогами.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:
6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.
6.1.2. *для лиц с нарушениями слуха:* в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.
6.1.3. *для лиц с нарушениями опорно-двигательного аппарата:* в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.