Программа учебной дисциплины
«Извлечение и анализ интернет-данных »

Утверждена
Академическим советом ОП
Протокол №_____ от _____._____.20_____

Разработчики	Голубев Илья, Департамент больших данных и информационного поиска
Число кредитов	3
Контактная работа (час.)	32
Самостоятельная работа (час.)	82
Курс, Образовательная программа	2-4 (Б), Экономика и статистика
Формат изучения дисциплины	Без использования онлайн курса

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. Ознакомление студентов с основными способами извлечения информации из интернета и эффективного анализа этой информации
2. Формирование у студентов практических навыков анализа и извлечения данных и работы с ними

Планируемые результаты обучения (ПРО):

1. 1) Знать основные языковые конструкции и типы данных языка python
2. 2) Владение инструментарием pandas: уметь работать с табличными данными средствами python
3. 3) Знание основных типов графиков и инструментов визуализации для python. Уметь изобразить гистограмму, диаграмму рассеяния, поточечный график. Уметь добавлять описание графика
4. 4) Понимать внутренню структуру форматов xml/json/html. Уметь обрабатывать файлы таких форматов с помощью модулей python: json, BeautifulSoup и др.
5. 5) Владеть инструментами python для доступа к web. Библиотека request.
6. 6) Иметь представление о модулях python для статистического анализа и машинного обученя
7. 7) Научиться пользоваться документацией языка и его библиотек

Пререквизиты:

1. МООС «Введение в Питон»
2. Основы теории вероятностей и статистики
3. Знаниями математики в объеме программы средней

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах** лк см onl/cp	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
Основы анализа данных в python	4 2 15	№: 1, 2, 7.	
Визуализация данных в python: библиотеки matplotlib, seaborn, plotly . Продвинутые инструменты для анализа данных	4 2 15	№: 3, 7.	
Парсинг открытых данных в различных форматах (xml/json/html)	2 4 15	№: 4, 7.	
Основы машинного обучения и практика применения	2 4 15	№: 6, 7.	
Извлечение и последующий анализ данных с ресурсов Википедия, Яндекс.Погода, Instagram	4 6 20	№: 5, 7.	
Часов по видам учебных занятий:	16 18 80		
Итого часов:	114		

Содержание разделов дисциплины:

1. **Основы анализа данных в python**
   Повторение основных функций и объектов языка Python. Обзор библиотек numpy, pandas на основе данных из соревнований платформы kaggle.com.
2. **Визуализация данных в python: библиотеки matplotlib, seaborn, plotly . Продвинутые инструменты для анализа данных**
   Введение в визуальный анализ данных. Построение графиков, гистограмм, тепловых карт. Знакомство с порталом Открытых данных.

3. **Парсинг открытых данных в различных форматах (xml/json/html)**
   Изучение языков и библиотек для работы с xml/json/html
4. **Основы машинного обучения и практика применения**
   Основные термины, понятия и алгоритмы машинного обучения. Обсуждение самых популярных систем, основанных на машинном обучении (распознавание изображений, поиск, диалоговые системы). Алгоритмы машинного обучения: линейная и логистическая регрессии, градиентный бустинг и нейронные сети.
5. **Извлечение и последующий анализ данных с ресурсов Википедия, Яндекс.Погода, Instagram**
   Извлечение данных с перечисленных ресурсов, их последующий анализ и обработка. Визуализация полученных данных.

3. Оценивание

- **СЕМ1**, Не блокирующее, Работа на семинаре
  Основная функциональность Питона. Задачи в классе
- **ДЗ1**, Не блокирующее, Домашнее задание
  Основная функциональность Питона. Самостоятельная работа
- **СЕМ2**, Не блокирующее, Работа на семинаре
  Работа с текстами. Задачи в классе
- **ДЗ2**, Не блокирующее, Домашнее задание
  Работа с текстами. Самостоятельная работа
- **СЕМ3**, Не блокирующее, Работа на семинаре
  Извлечение данных с ресурсов. Задачи в классе
- **ДЗ3**, Не блокирующее, Домашнее задание
  Извлечение данных с ресурсов. Самостоятельная работа
- **СЕМ4**, Не блокирующее, Работа на семинаре
  Продвинутое извлечение данных с ресурсов. Задачи в классе
- **ДЗ4**, Не блокирующее, Домашнее задание
  Продвинутое извлечение данных с ресурсов. Самостоятельная работа
- **СЕМ5**, Не блокирующее, Работа на семинаре
  Извлечение и визулизация данных. Задачи в классе
- **ДЗ5**, Не блокирующее, Домашнее задание
  Извлечение и визулизация данных. Самостоятельная работа
- **ЭКЗ1**, Не блокирующее, Экзамен (письменный)
  Письменный экзамен: задачи и извлечение из популярных ресурсов и последующий анализ данных

Формула округления: Стандартное арифметическое округление

Вид формулы оценивания: Линейная
Формула оценивания:

Окончательная оценка = Округление(0.35 * ДЗ + 0.35 * Семинары + 0.3 * Экзамен)

4. Примеры оценочных средств

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	Документация языка python
2	Data Analysis in Python
3	Документация Pandas
4	Документация matplotlib
5	Примеры визуализации в python
6	Документация BeautifulSoup
7	*Мария Мансурова* Web Scraping с помощью python
8	Документация scikit-learn
9	Статистические методы в python

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	Anaconda - это бесплатный дистрибутив языка программирования Python для научных вычислений с открытым исходным кодом,	Бесплатно с https://www.anaconda.com/

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/
	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.

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.