Программа учебной дисциплины

«Автоматизированный сбор больших данных в экономико-социологических исследованиях»

Утверждена

Академическим советом ООП

Протокол № 03 от «05» июня 2018 г.

Автор	Управителев Филипп Александрович
Число кредитов	4
Контактная работа (час.)	60
Самостоятельная работа (час.)	92
Курс	1
Формат изучения дисциплины	без использования онлайн курса

ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Автоматизированный сбор больших данных в экономико-социологических исследованиях» получение представления о роли данных в современном мире и формирование базовых навыков работы с большими данными.

В результате освоения дисциплины студент должен:

- уметь формулировать цели и задачи исследования, а также соответствующие им
- гипотезы;
- уметь искать статистические данные, необходимые для анализа и проводить с ними предварительную работу по оценки качества данных и устранению ошибок и расхождений;
- уметь подбирать статистические методы под задачи исследования, которые возможно реализовать на выбранных данных;
- уметь выполнять статистический анализ данных в среде R и корректно его интерпретировать;
- уметь излагать результаты исследования.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать математику в объеме средней школы;
- уметь самостоятельно формулировать цели, ставить конкретные задачи научных исследований в фундаментальных и прикладных областях.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- Научно-исследовательский семинар
- Методы анализа больших данных в исследованиях поведения потребителей

СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в большие данные - идеи, технологии, методы и области применения

Развитие технологий. Web2.0, удешевление технологий хранения, облачные технологии, интернет вещей, quantified self. Многообразие доступных данных. Тренды на открытую науку и предоставление данных в открытый доступ. Data-driven подход. Развитие машинного обучения и прочих методов анализа данных.

Основы языка R. История и развитие языка, основная сфера применения. Введение в

R. Установка, рабочие панели RStudio, базовые типы и структуры данных. Циклы и векторизованные функции. Работа с большими данными в R. Пакет data.table. Особенности синтаксиса data.table. Агрегация данных, слияние таблиц, прочие трансформации. Создание новых колонок. Фильтрация по строкам.

Тема 2. Виды источников данных

Этапы ETL. Структурированные и неструктурированные типы данных. Основные форматы файлов - txt, csv, xls, sav. Структура файлов. Виды разделителей, символы окончания строки. Проблема кодировок и различия операционных систем. SQL-базы данных. Удаленные базы данных (API). Неструктурированные данные - json, xml. NoSQL-базы данных. Сохранение или запись файлов, представление в внешних веб-приложения.

Импорт текстовых файлов. Импорт спец.форматов - xlsx, sav. Ошибки при импорте.

Тема 3. Методы сбора удаленных данных

Сайты как источник данных. Парсинг сайтов. HTML, XPath, DOM-разметка. Пакет rvest.

Подключение и импорт данных из базы данных. Подключение к API. Простейшие парсеры.

Описательные статистики. Меры центральных тенденций. Выбросы, пропущенные значения. Обработка пропущенных значений.

Тема 4. Представление результатов исследования

Задачи визуализации данных. Статичные графики, интерактивные визуализации, инфографика. Виды графиков - описательные, статистические, геокарты, многомерные графики.

Принципы визуальной презентации данных. Ошибки в использовании линейных графиков, гистограмм, круговых и объемных диаграмм. Палитры для графиков.

Пакет ggplot. Принцип слоев. Основные виды графиков в ggplot. Кастомизация графиков - цвета, оси, аннотации и тексты. Комбинированные графики, фасеты.

Интерактивные графики. Пакет plotly. Основные виды графиков в plotly. Структура графиков plotly в json-записи. Добавление слоев. Ховеры, комбинированные графики, двойные оси. Публикация графиков. Импорт ggplot-объектов.

Хранение проекта, структура папок. Основные элементы проекта. Репродуцируемые отчеты. Язык разметки markdown. Структура заголовков, чанки. Форматирование, таблицы и проч. Простейшие отчеты в markdown.

Тема 5. Анализ данных на естественном языке

Основные идеи NLP. Мешок слов, n-граммы, тематическое моделирование, word2vec, LSTM-модели. Анализ тональности. Корпусы русского языка.

Стемминг и нормализация. Облако слов. Пакеты lda и text2vec. Тематическое моделирование.

Тема 6. Машинное обучение. Обучение без учителя

Проверка гипотез на больших данных. Задачи машинного обучения. Обучение без учителя. Иерархический кластерный анализ. K-means. Метод главных компонент (PCA). Проблемы интерпретации и генерализации.

Кластеризация. Снижение размерности.

Тема 7. Машинное обучение. Обучение с учителем

Обучение с учителем, регрессионные модели и случайные леса. Этапы построения модели - препроцессинг, выбор и конструирование пространства признаков (фичей), выбор метода и метрик. Обучающая и тестовая выборки. Ансамбли моделей.

Линейные регрессионные модели. Предсказание классов. Случайные лес.

ОЦЕНИВАНИЕ

Итоговая оценка по дисциплине складывается из накопленных оценок за домашние задания и проектную работу. В ходе курса студенты могут получить 30 баллов за три домашних задания (по 10 за каждое) и 20 баллов за итоговую проектную работу.

Для получения оценки по 10-ти балльной шкале сумма набранных баллов делится на 5 и округляется арифметически.

В случае, если домашнее задание сдано позже установленного срока (но не более чем на 7 дней), оценка снижается на 1 балл. В более поздние сроки задания не принимаются. Текущие домашние задания выдаются и принимаются по мере прохождения программы, итоговый проект принимается не позднее, чем за неделю до начала сессии третьего модуля. Оценки за курс выставляются в течение сессии третьего модуля.

ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

1. Импортируйте *.csv-файл, с учетом нестандартных разделителей

2. Импортируйте *.sav-файл с сохранением меток
3. Импортируйте таблицу данных из удаленной PostgreSQL-базы
4. Визуализуйте динамику цен на недвижимость с линией сглаженного среднего
5. Визуализируйте интерактивную тепловую карту количества покупок в онлайн- магазине по часам и по районам Москвы
6. Сегментируйте страны по значениям ВВП, аргументируйте выбор количества сегментов

РЕСУРСЫ

1 Основная литература

1. Tattar, Prabhanjan N., et al. A Course in Statistics with R, John Wiley & Sons, Incorporated, 2016. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4452971

Thomas, Seemon. Basic Statistics, Alpha Science Internation, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=5190782

2 Дополнительная литература

Racine, J. S.Nonparametric and semiparametric methods in R, 2009, https://www.emeraldinsight.com/doi/full/10.1108/S0731-9053%282009%290000025014

Rohatgi, Vijay K., and A. K. Md. Ehsanes Saleh. Introduction to Probability and Statistics, Wiley, 1994. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=818930.

3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	RStudio	Свободно распространяемое лицензионное соглашение
2.	R	Свободно распространяемое лицензионное соглашение
3.	Microsoft Windows 7	Из внутренней сети университета (договор)
4.	Microsoft Office Professional Plus 2010	Из внутренней сети университета (договор)

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

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Электронно-библиотечные ресурсы НИУ	URL: https://library.hse.ru/e-resources

	ВШЭ	
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/

5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет;
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для самостоятельных занятий по дисциплине оснащены ноутбуками, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.