БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____ О.И.Чуприс
«12» 04 2019 г.

Регистрационный № УД-6987/уч.

ЯЗЫК ПРОГРАММИРОВАНИЯ R В СОЦИАЛЬНЫХ НАУКАХ

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:

1-23 81 08 Медиакоммуникации

2019 г.

Учебная программа составлена на основе:
ОСВО 1-23 81 08-2018 и учебного плана Е23-316/уч. от 11.05.2018

СОСТАВИТЕЛЬ:
М. С, Фабрикант, доцент кафедры психологии факультета философии и социальных наук Белорусского государственного университета, кандидат психологических наук

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой психологии
(протокол № 12 от 27.06.2019);

Научно-методическим Советом БГУ

(протокол № 5 от 28.06.2019)

Заведующий кафедрой
д. психол. н., профессор _____ И. А. Фурманов

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Язык программирования R – специализированный язык, предназначенный для автоматизированного компьютерного анализа данных, который является одним из основных инструментов современного аналитика и исследователя данных. В социальных науках R позволяет использовать разнообразные методы статистической обработки данных в различных форматах и различного происхождения – от структурированных опросных данных до неструктурированных текстов, сгенерированных естественным образом.

«Язык программирования R в социальных науках» – практико-ориентированная дисциплина, направленная на использование современных способов автоматического компьютерного анализа данных в фундаментальных и прикладных исследованиях, проводимых социальными учеными. Освоение анализа данных на R существенно повышает конкурентоспособность специалиста на рынке труда.

Цели и задачи учебной дисциплины

Цель учебной дисциплины – подготовка специалиста в области социальных наук, владеющего современными техниками написания программного кода для статистического анализа данных методами, наиболее часто используемыми в социальных науках в настоящее время.

Задачи учебной дисциплины:

1. изучить основные принципы, приемы и техники программирования на языке R;
2. изучить способы реализации при помощи языка R наиболее часто используемых современных методов статистического анализа данных в социальных науках;
3. изучить типичные способы применения, возможности и ограничения программных возможностей R в социальных науках.

Место учебной дисциплины в системе подготовки магистра.

Учебная дисциплина «Язык программирования R в социальных науках» относится к циклу специальных дисциплин (компонент учреждения образования) профиля «Медиаисследования и социальная аналитика» модуля «Статистический анализ данных в медиаисследованиях».

Учебная дисциплина «Язык программирования R в социальных науках» читается в третьем семестре и тесно связана с другими дисциплинами учебного плана, такими как «Анализ данных в социальных науках», «Многомерный анализ данных», «Компьютерный анализ данных», «Основы выборочного метода» и «Онлайн-исследования»

Требования к компетенциям

Освоение учебной дисциплины «Язык программирования R в социальных науках» должно обеспечить формирование универсальных, углубленных профессиональных и специализированных компетенции:

универсальные компетенции:

УК-7. Уметь применять междисциплинарные научные знания для постановки и решения производственных задач.

углубленные профессиональные компетенции:

УПК-2. Владеть современными методами сбора, обработки, анализа, представления и распространения информации с использованием новейших информационно-коммуникационных технологий.

специализированные компетенции:

2СК-9. Быть способным использовать язык программирования R в социальных исследованиях

В результате освоения учебной дисциплины студент должен:

знать:

- основные принципы программирования на языке R;
- основные методы статистического анализа дынных в социальных науках и соответствующие им функции и программные пакеты в R;
- основные способы визуализации данных в R;

уметь:

- использовать функции и программные пакеты языка R ля статистического анализа данных;
- интерпретировать результаты анализа данных в R для различных статистических методов;
- кастомизировать и создавать понятный для других специалистов и пригодный для многократного использования программный код на языке R;

владеть:

- основами функционального и объектно-ориентированного программирования на языке R;
- навыками выбора подходящего метода анализа данных для решения конкретных исследовательских задач и подбора соответствующих им программных функций и пакетов в R;
- навыками построения проверки, оценки качества и корректировки статистических моделей посредством R.

Структура учебной дисциплины

Дисциплина изучается в 3 семестре. Всего на изучение учебной дисциплины «Язык программирования R в социальных науках» отведено 108 часов, в том числе 36 аудиторных часов, из них: лекции – 18часов, семинарские занятия – 18 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Общие принципы анализа данных в R

Основные принципы программирования в R и история его создания. Интерактивные среды для написания программного кода. Способы инсталляции и активации дополнительных программных пакетов. Структура описаний программных пакетов. Форум StackOverflow.

Типы данных в R. Чтение баз данных в различных форматах. Основы функционального и объект-ориентированного программирования в R. Преобразование типов шкал.

Тема 2. Вычисление основных описательных статистик в R

Расчет одномерных частотных распределений, накопленных частот и процентов и таблиц сопряженности. Меры центральной тенденции: арифметическое среднее, мода, медиана. Меры разброса данных: минимум, максимум, размах, дисперсия, стандартное и среднее квадратическое отклонение. Квартильная и другие группировки. Коэффициенты корреляции. Доверительный интервал и уровень значимости. Написание функций для расчета описательных статистик и использование имеющихся функций. Кастомизация формата вывода результатов.

Тема 3. Визуализация данных

Основные принципы визуализации данных. Графики на частотных распределениях: круговая диаграмма, столбиковая диаграмма, гистограмма. Графики на квартильных группировках: коробчатая и скрипичная диаграммы. Визуализация корреляций: график рассеяния, добавление графика линейной функции, отображение доверительного интервала. Кастомизация графиков и их сохранение в различных форматах. Основы работы с программным пакетом ggplot2.

Тема 4. Машинное обучение с супервизией: регрессионный анализ

Понятие машинного обучения с супервизией. Маркировка данных. Понятие зависимых и независимых переменных. Общая логика построения и проверки моделей в предиктивной аналитике. Виды регрессионного анализа.

Линейный регрессионный анализа. Нестандартизированный и стандартизированный регрессионные коэффициенты в линейной регрессии. Логистическая регрессия: бинарная, мультиноминальная, ординальная. Коэффициенты логистической регрессии: отношение шансов и логит. Генерализованная регрессионная модель.

Параметры оценки качества модели. Процент объясненной дисперсии, байесовские критерии, критерий наибольшего правдоподобия. Дифференцированная оценка точности предсказания для различных категорий значений зависимой переменной.

Тема 5. Машинное обучение с супервизией: классификация

Логика построения классификаторных моделей на маркированных данных. Классификация переменных: конфирматорный факторный анализ при помощи программного пакета lavaan. Интерпретация факторных нагрузок и параметров оценки качества модели. Корректировка модели. Понятие и уровни инвариантности в мульгрупповом конфирматорном факторном анализе. Основы моделирования структурными уравнениями.

Классификация кейсов: метод К-ближайших соседей. Варьирование параметров, проверка аи оптимизация модели.

Сравнительная оценка значимости дифференцирующих критериев классификации: деревья решений и случайные леса.

Тема 6. Машинное обучение без супервизии: кластеризация

Понятие машинного обучения без супервизии. Маркировка недифференцированных данных. Логика и область применения кластеризации. Виды кластерного анализа: К-средних и иерархический кластерный анализ. Построение и чтение дендрограммы. Оценка качества модели и ее использование в дальнейшем анализе данных.

Тема 7. Основы обработки неструктурированных данных

Общие принципы статистического анализа текстов. Определение единиц анализа. Подготовка данных: вебскрейпинг, объединение и разметка массивов. Наивный Байес как основной алгоритм распознавания неструктурированных текстовых данных: причины несовершенства и пути усовершенствования. Основы анализа визуальных данных: компьютерное зрение и глубокое обучение. Основы моделирования нейронных сетей.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования

Номер раздела,	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	Общие принципы анализа данных в R	2		2				коллоквиум
2	Вычисление основных описательных статистик в R	2		2				решение задач
3	Визуализация данных	2		2				решение задач
4	Машинное обучение с супервизией: регрессионный анализ	4		4				проект
5	Машинное обучение с супервизией: классификация	4		4				проект
6	Машинное обучение без супервизии: кластеризация	2		2				проект
7	Основы обработки неструктурированных данных	2		2				решение задач

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Зарядов, И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. – М.: Изд-во РУДНБ, 2010. – 207 с.

3. Наглядная статистика. Используем R! [А. Б. Шипунов и др.]. – Москва: ДМК Пресс, 2014. – 298 с.

4. Мастицкий, С.Э. Статистический анализ и визуализация данных с помощью R / С.Э. Мастицкий, В. К. Шитиков. М.: ДМК, 2015. – 401 с.

Перечень дополнительной литературы

1. Agresti, A. An introduction to categorical data analysis / A. Agresti. – London: Wiley, 2018. – 400 p.

2. Alvarez, R. M. Computational social science / R. M. Alvarez. – Cambridge: Cambridge University Press, 2016. – 338 p.

3. Faraway, J. J. Linear models with R / J. J. Faraway. – NYC: Chapman and Hall/CRC, 2016. – 240 p.

4. Fieller, N. Basics of matrix algebra for statistics with R / N. Fieller. – NYC: Chapman and Hall/CRC, 2018. – 248 p.

5. Heeringa, S.G. Applied survey data analysis / S.G. Heeringa, B.T. West, P.A.. Berglund, – NYC: Chapman and Hall/CRC, 2017. – 487 p.

6. Murrell, P. R graphics / P. Murrell. NYC: Chapman and Hall/CRC, 2016. – 423 p.

7. Wickham, H. ggplot2: elegant graphics for data analysis / H. Wickham. – NYC: Springer, 2016. – 213 p.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

- Коллоквиум
- Решение задач
- Проект

Примерная тематика семинарских занятий

Семинар № 1. Общие принципы анализа данных в R

Семинар № 2. Вычисление основных описательных статистик в R

Семинар № 3. Визуализация данных

Семинар № 4. Машинное обучение с супервизией: регрессионный анализ

Семинар № 5. Машинное обучение с супервизией: классификация
Семинар № 6. Машинное обучение без супервизии: кластеризация
Семинар № 7. Основы обработки неструктурированных данных

Описание инновационных подходов и методов к преподаванию учебной дисциплины (эвристический, проективный, практико-ориентированный)

При организации образовательного процесса используется метод проектного обучения, который предполагает:

- способ организации учебной деятельности студентов, развивающий актуальные для учебной и профессиональной деятельности навыки планирования, самоорганизации, сотрудничества и предполагающий создание собственного продукта;
- приобретение навыков для решения исследовательских, творческих, социальных, предпринимательских и коммуникационных задач.

Также при организации образовательного процесса используется ***практико-ориентированный подход,*** который предполагает***:***

- освоение содержание образования через решения практических задач;
- приобретение навыков эффективного выполнения разных видов профессиональной деятельности;
- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
- использованию процедур, способов оценивания, фиксирующих сформированность профессиональных компетенций.

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы студентов по учебной дисциплине «Язык программирования R в социальных науках» в распоряжение учащихся предоставляются учебные и учебно-методические материалы:

- учебная программа;
- учебные издания для теоретического изучения дисциплины;
- методические указания к семинарским занятиям;
- методические рекомендации для выполнения коллоквиума;
- образцы программного кода;
- публикации с примерами использования R в социальных науках;
- список основной литературы и дополнительной;
- перечень информационных ресурсов.

Методика формирования итоговой оценки

Итоговая оценка формируется на основе:

1. Правил проведения аттестации студентов (Постановление Министерства образования Республики Беларусь №53 от 29.05.2012 г.);
2. Положения о рейтинговой системе оценки знаний по дисциплине в БГУ (приказ ректора БГУ от 18.08.2015 г. №382-ОД);
3. Критериев оценки знаний студентов (письмо Министерства образования Республики Беларусь от 22.12.2003 г.)

Примерный перечень вопросов к зачету

1. Основные принципы программирования в R и история его создания. Интерактивные среды для написания программного кода.
2. Способы инсталляции и активации дополнительных программных пакетов. Структура описаний программных пакетов. Форум StackOverflow.
3. Типы данных в R. Чтение баз данных в различных форматах.
4. Основы функционального и объект-ориентированного программирования в R. Преобразование типов шкал.
5. Расчет одномерных частотных распределений, накопленных частот и процентов и таблиц сопряженности.
6. Меры центральной тенденции: арифметическое среднее, мода, медиана.
7. Меры разброса данных: минимум, максимум, размах, дисперсия, стандартное и среднее квадратическое отклонение.
8. Квартильная и другие группировки.
9. Коэффициенты корреляции.
10. Доверительный интервал и уровень значимости.
11. Написание функций для расчета описательных статистик и использование имеющихся функций.
12. Кастомизация формата вывода результатов.
13. Основные принципы визуализации данных.
14. Графики на частотных распределениях: круговая диаграмма, столбиковая диаграмма, гистограмма.
15. Графики на квартильных группировках: коробчатая и скрипичная диаграммы. Визуализация корреляций: график рассеяния, добавление графика линейной функции, отображение доверительного интервала.
16. Кастомизация графиков и их сохранение в различных форматах.
17. Основы работы с программным пакетом ggplot2.
18. Понятие машинного обучения с супервизией. Маркировка данных.
19. Понятие зависимых и независимых переменных. Общая логика построения и проверки моделей в предиктивной аналитике.
20. Виды регрессионного анализа.
21. Линейный регрессионный анализа.

22. Нестандартизированный и стандартизированный регрессионные коэффициенты в линейной регрессии.
23. Логистическая регрессия: бинарная, мультиноминальная, ординальная.
24. Коэффициенты логистической регрессии: отношение шансов и логит.
25. Генерализованная регрессионная модель.
26. Параметры оценки качества модели. Процент объясненной дисперсии, байесовские критерии, критерий наибольшего правдоподобия.
27. Дифференцированная оценка точности предсказания для различных категорий значений зависимой переменной.
28. Логика построения классификаторных моделей на маркированных данных.
29. Классификация переменных: конфирматорный факторный анализ при помощи программного пакета lavaan.
30. Интерпретация факторных нагрузок и параметров оценки качества модели. Корректировка модели.
31. Понятие и уровни инвариантности в мульгрупповом конфирматорном факторном анализе.
32. Основы моделирования структурными уравнениями.
33. Классификация кейсов: метод К-ближайших соседей. Варьирование параметров, проверка аи оптимизация модели.
34. Сравнительная оценка значимости дифференцирующих критериев классификации: деревья решений и случайные леса.
35. Понятие машинного обучения без супервизии. Маркировка недифференцированных данных.
36. Логика и область применения кластеризации.
37. Виды кластерного анализа: К-средних и иерархический кластерный анализ.
38. Построение и чтение дендрограммы.
39. Оценка качества кластерной модели и ее использование в дальнейшем анализе данных.
40. Общие принципы статистического анализа текстов. Определение единиц анализа.
41. Подготовка данных для анализа текста: вебскрейпинг, объединение и разметка массивов.
42. Наивный Байес как основной алгоритм распознавания неструктурированных текстовых данных: причины несовершенства и пути усовершенствования.
43. Основы анализа визуальных данных: компьютерное зрение и глубокое обучение.
44. Основы моделирования нейронных сетей.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
1. Базы данных в социальных науках	Кафедра социальной коммуникации	Нет изменений	Изменений не требуется, протокол № 6 от 14 03.2019
2.Инфографика	Кафедра психологии	Нет изменений	Изменений не требуется, протокол № 6 от 14 03.2019

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры психологии (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой
Д. психол. н., профессор _____ И. А. Фурманов

УТВЕРЖДАЮ
Декан факультета
К. и. н., доцент _____ В. Ф. Гигин