Программа учебной дисциплины «Введение в науку о данных»

Утверждена
Академическим советом ОП
Протокол № 15 от «28» июня 2018 г.

Автор	Г. А. Мороз, старший преподаватель, Школа лингвистики ФГН
Число кредитов	8
Контактная работа (час.)	160
Самостоятельная работа (час.)	224
Курс	1
Формат изучения дисциплины	без использования онлайн-курса

1. Цель, результаты и пререквизиты освоения дисциплины

Целями освоения дисциплины «Введение в науку о данных» является знакомство:
с основами работы в R и RStudio
с основными типами данных (таблицы, тексты, изображение с текстом);
с основными методами сбора, обработки и трансформации данных;
с основными методами визуализации и представления данных;
с основными методами статистического анализа;
с основными методами регрессионного анализа;
с основными методами кластерного анализа;
с основными методами уменьшения размерностей;
с основными методами сетевого анализа;
с основными методами байесовского анализа данных;

В результате освоения дисциплины студент должен:

знать

- особенности работы R, основные особенности анализа различных типов данных;
- познакомиться с основами методами регрессионного, кластерного и сетевого анализа, методами уменьшения размерностей и байесовского анализа данных

уметь

- подготавливать данные из разных типов источников;
- преобразовывать и визуализировать данные;
- применять методы регрессионного, кластерного и сетевого анализа, методами уменьшения размерностей и байесовского анализа данных

владеть

- базовыми навыками самостоятельного анализа данных, а также критической интерпретации анализа данных, представленной в научных работах;

У курса нет пререквизитов.

Формат изучения: лекции, семинары, без использования онлайн курсов.

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах**	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
Тема 1. Основы R	лк 2 см 3 ср 40	Например: написать функцию, которая считает факториал.	Один из вопросов из домашней работы
Тема 2. Сбор и обработка данных	лк 4 см 10 ср 41	Например: напишите программу, которая посчитает долю слов, в заголовках новостей магистерской программы «Цифровые методы в гуманитарных науках».	Один из вопросов из домашней работы
Тема 3. Визуализация и представление данных	лк 2 см 6 ср 24	Например: постройте столбчатую диаграмму 7 наиболее частотных слов из заголовков новостей магистерской программы «Цифровые методы в гуманитарных науках».	Один из вопросов из домашней работы
Тема 4. Основы статистики	лк 4 см 10 ср 40	Например: посчитайте долю использования слова «not» в корпусе обычных писем и спама. Является ли наблюдаемая разница средних статистически значимой?	Один из вопросов из домашней работы
Тема 5. Регрессионный анализ	лк 4 см 10 ср 40	Например: постройте пуассоновскую регрессию, предсказывающую количество наград, которые получат студенты на основе типа программы и оценки их экзамена по математике.	Один из вопросов из домашней работы
Тема 6. Кластерный анализ	лк 1 см 2 ср 7	Например:	Один из вопросов из домашней работы
Тема 7. Методы уменьшения размерности	лк 1 см 2 ср 12	Например: Используйте PCA для анализа датасета с частотностями прилагательных поэтов серебряного века. Какие кластеры поэтов можно обнаружить?	Один из вопросов из домашней работы
Тема 8. Сетевой анализ	лк 1 см 2 ср 7	Например: провести разграничение между понятиями норма и узус, язык и речь.	Один из вопросов из домашней работы
Тема 9. Байесовские методы	лк 5 см 13	Например: Посчитайте значение правдоподобия	Один из вопросов из домашней работы

	ср 52	модели $\mathcal{N}(\mu{=}910, var{=}150)$ для встроенного датасета Nile.	
Часов по видам учебных занятий:	лк 24		
	см 56		
	ср 224		
Итого часов:	304		

Тема 1. Основы R:
Базовые объекты, функции, пакеты. Написание собственных функций и сложные циклы

Тема 2. Сбор и обработка данных:
Трансформация данных: tidyverse, dplyr. Работа со строками. Работа с текстами: tidytext, udpipe. Сбор данных из интернета: rvest. OCR

Тема 3. Визуализация и представление данных:
Визуализация данных: ggplot2. Представление данных: rmarkdown, shiny. Работа с картографическими данными. Визуализация данных: графы, санки-плот.

Тема 4. Основы статистики:
Описательная статистика, моменты, z-преобразование. Центральная предельная теорема. Доверительные интервалы, T-test, χ^2, Fisher-test. Критерии согласия. Симуляционная статистика.

Тема 5. Регрессионный анализ:
Корреляция и простая линейная регрессия. Множественная регрессия, link-functions. Логистическая и мультиномиальная регрессия. GAM. Ограничение на применение регрессии. Модели со смешанными эффектами.

Тема 6. Кластерный анализ:
Методы кластеризации: метрики расстояний, k-means, иерархические кластеризации. Визаулизация деревьев.

Тема 7. Методы уменьшения размерности:
PCA, CA, MCA, MDS, t-SNE

Тема 8. Сетевой анализ:
Основные метрики сетей. Методы визуализации сетей.

Тема 9. Байесовские методы:
Работа с распределениями. Байесовский статистический вывод. Байесовский доверительный интервал. Байесовский фактор. Эмпирическая байесовская оценка. Байесовкие A/B тесты. Байесовская регрессия. Введение в MC и MCMC. Пакет brms

3. Оценивание

Итоговая оценка за курс состоит из оценок за самостоятельные работы по следующей формуле:

$$O_{итоговый} = \frac{\sum_{i=1}^{12} 10 \times m_{i,k}}{12},$$

где $m_{i,k}$ – доля студентов, выполнившая задание номер i хуже, чем студент k.

Оценки выставляются по 10-балльной шкале. Способ округления оценок: арифметический.

Оценка 10 ставится за абсолютно верный ответ, содержащий элементы нетривиального подхода к анализу материала.
Оценка 9 ставится за абсолютно верный ответ, не обладающий нетривиальными особенностями.
Оценка 8 ставится за абсолютно верный ответ с незначительными погрешностями при условии их самостоятельного исправления в процессе диалога с преподавателем.

Ни одна из оценок не является блокирующей.

Все элементы контроля подлежат пересдаче в виде 2-ух часовой работы, во время которой можно пользоваться любыми материалами. Время проведения устанавливается факультетом гуманитарных наук. Тематический состав КИМ-ов для пересдач не отличается от тематического состава КИМ-ов текущего контроля и промежуточной аттестации.

4.Примеры оценочных средств

Уже приведены в разделе 2.

5. Ресурсы

5.1 Рекомендуемая основная литература

Базовый учебник по курсу отсутствует.

5.2 Рекомендуемая дополнительная литература

№ п/п	Наименование
1.	Xie, Y. Dynamic documents with R and knitr / Y. Xie. – Boca Raton; London; New York: CRC Press, 2014. – 190 с. – (The R series) . – На англ. яз. - ISBN 978-1-482-20353-0.
2.	Spector, P. Data manipulation with R / P. Spector. – New York: Springer, 2008. – 152 с. – (Use R!). – На англ. яз. - ISBN 978-0-387-74730-9.
3.	Wickham, H. ggplot2: elegant graphics for data analysis / H. Wickham. – Dordrecht: Springer, 2009. – 212 с. – (Use R!). – На англ. яз. - ISBN 978-0-387-98140-6.
4.	Wickham, H. Advanced R / H. Wickham. – Boca Raton [etc.]: CRC Press, 2014. – 456 с. – (Chapman & Hall/CRC. The R Series) . – На англ. яз. - ISBN 978-1-466-58696-3.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	R	*Распространяется бесплатно*
2.	RStudio	*Распространяется бесплатно*
3.	Git	*Распространяется бесплатно*

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

№ п/п	Наименование	Условия доступа
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Github	https://github.com/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных и семинарских занятий по дисциплине оснащены ПЭВМ по числу студентов 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:

i. *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.

ii. *для лиц с нарушениями слуха*: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

iii. *для лиц с нарушениями опорно-двигательного аппарата*: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.