Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего профессионального образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет Санкт-Петербургская школа социальных и гуманитарных наук

Программа дисциплины «Введение в науку о данных»

для направления 39.03.01 “Социология”
подготовки бакалавра

Автор программы:
Мусабиров И.Л. ilya@musabirov.info, imusabirov@hse.ru
Окопный П.В.
Иванюшина В.А., канд.биол.наук, доцент
Сироткин А.В., канд физ.-мат. наук, доцент

Согласована методистом ОСУП

«_____»_____2014 г. _____ Т.Г. Ефимова

Утверждена академическим руководителем ОП “Социология”

«___»_____2015 г. _____ Д.А.Александров

Санкт-Петербург, 2015

Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности. Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления подготовки 39.03.01 «Социология» подготовки бакалавра, изучающих дисциплину «Введение в науку о данных».

Программа разработана в соответствии с:

- Образовательным стандартом Федерального государственного автономного образовательного учреждения высшего профессионального образования «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ – ВЫСШАЯ ШКОЛА ЭКОНОМИКИ» в редакции 2014г.
- Рабочим учебным планом университета по направлению подготовки бакалавра 39.03.01 «Социология»для студентов 1 курса, утверждённым в 2014 году.
- Образовательной программой НИУ ВШЭ по направлению подготовки бакалавра 39.03.01 «Социология».

Цели освоения дисциплины

Цель освоения дисциплины «Введение в науку о данных» — познакомить студентов с проблематикой и методами Data Science, дать представление об их применении в научных исследованиях и на практике.

Компетенции обучающегося, формируемые в результате освоения дисциплины

Компетенция	Код по НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
владение основными методами, способами и средствами получения, хранения, переработки информации, навыки работы с компьютером как средством управления информацией	ОК-13	Владеет навыками программирования на языке R, знает способы автоматизированной обработки, очистки данных, владеет навыками работы с системами управления базами данных	Практические и самостоятельные занятия по написанию программ для сбора и обработки информации
способность работать с информацией в глобальных компьютерных сетях	ОК-14	Владеет навыками написания программ для автоматизированного сбора информации из различных источников в глобальных компьютерных сетях	Самостоятельная и проектно-групповая деятельность по анализу результатов проведенных исследований

способность самостоятельно формулировать цели, ставить конкретные задачи научных исследований в различных областях социологии и решать их с помощью современных исследовательских методов и информационных технологий	ПК-2	Определяет актуальные для исследования проблемы, формулирует цели задачи и исследования, выбирает методы сбора данных и стратегии исследования	Семинарские занятия по обсуждению плана и этапов исследования, формирование исследовательских групп, анализ и обсуждение проведенных исследований
способность и готовность участвовать в составлении и оформлении научно-технической документации, научных отчетов, представлять результаты исследовательской работы с учётом особенностей потенциальной аудитории	ПК-3	Может подготовить сообщение в разной форме (презентация, постерный доклад, публикация) о результатах исследоания	Семинарские занятия, домашнее задание, участие в студенческих научных конференциях
умение обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций	ПК-8	Владеет навыками преобразования данных исследования в экспертные решения и рекомендации в определенной области	Практические и семинарские занятия, домашняя работа, проектная работа

Место дисциплины в структуре образовательной программы

Настоящий факультативный курс рассчитан на студентов-социологов, знакомых с курсом математики в объёме средней школы, прослушавших, или параллельно слушаюших курсы «Прикладное ПО», «Теория вероятностей и математическая статистика».

Настоящий курс подготавливает студентов к майнору Data Science, участию в проектной деятельности с применением цифровых методов исследований.

Приобретённые знания и умения будут полезны студентам в курсах «Анализ данных в социологии», «Компьютерные методы анализа текста», «Социология социальных сетей», научно-исследовательской и практической работе, включающей применение количественных и цифровых методов исследований.

Курс является прямым логическим продолженем дисциплины «Введение в моделирование в социальных науках» и предполагает взаимодействие со студентами-первокурсниками, изучающими данные курсы, в рамках проектных заданий.

Курс закладывает прочный фундамент для выполнения курсовых и выпускных квалификационных работ, практических проектов в рамках междисциплинарного направления Digital Social Science.

Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторных часов		Самостоят. работа
			Лекции	Практич. занятия	

1	Введение в Data Science	33	4	6	7
2	Современная инфраструктура обработки данных. Data Workflows	33	10	12	19
3	Введение в интеллектуальный анализ данных, сетей и текстов	33	10	12	19
4	Проектная организация цифровых исследований	53	8	10	35
Всего		152	32	40	80

Формы контроля знаний студентов

Тип контроля	Форма контроля	Модуль				Параметры
		Первый	Второй	Третий	Четвёртый	
Текущий	Контрольная работа			* *	*	[1]
Итоговый	Экзамен				*	[2]

Критерии оценки знаний, навыков

Критерии оценки заданий, связанных с написанием программного кода (в составе домашних и контрольных работ)

1. Полнота, качество и скорость выполнения задания, уровень навыка использования средств разработки и широта диапазона корректно применяемых языковых конструкций и паттернов.
2. Соответствие принципам *literate programming*(грамотного программирования): встроенность программного кода в контекст решения задачи, включающий текстовые описания, программногенерируемые иллюстрации и пр.
3. Грамотная организация сложного кода, корректное применение принципов и средств структурного и функционального программирования.

На практических занятиях акцент делается преимущественно на критерии 1 группы; поощряется самостоятельное знакомства с дополнительными свойствами и возможностями программных конструкций по рассматриваемой теме без выпадения из потока практического занятия; самостоятельное ведение средствами рабочей среды конспекта занятия.

При выполнении *самостоятельных заданий* с использованием дистанционного доступа к рабочей среде и **сквозных программных заданий** дополнительный акцент делается на соответствие 2 и 3 группам критериев (с постепенным ростом значения 3 группы по ходу курса).

[1] Контрольные работы 3 модуля предполагают выполнение вычислительных заданий по материалу модуля и ответов на вопросы по статьям, связанным с исследованиями в области CSS. Контрольная работа 4 модуля предполагает выполнение комплексного вычислительного задания, реплицируюшего реальное исследование в области CSS.

[2] Экзамен проходит в форме защиты проекта и ответов на дополнительные вопросы по применяемым методам.

Критерии оценки заданий, включающих элементы анализа данных (в составе домашних и контрольных работ)

- корректность применения методик анализа (в рамках знаний, полученных в курсе, смежных дисциплинах, домашнем чтении);
- уверенность использования языковых средств и структур данных, методов преобразования и агрегации данных в организации потока анализа данных, их ввода и вывода;
- уровень включённости процедур анализа данных в контекст решения содержательной задачи.

Порядок формирования оценок по дисциплине

Накопленная оценка за текущий контроль по модулю формируется следующим образом:

$$O_{\text{накопленная}} = (\frac{1}{2} \cdot (0.8 \cdot O_{\text{сам.раб}} + 0.2 \cdot O_{\text{аудит.}}) + \frac{1}{2} \cdot O_{\text{арифм.ср.по контрольным}})$$

.

Способ округления при выставлении накопленной оценки текущего контроля — арифметический.

Экзаменационная оценка в модуле 4 выставляется по итогам презентации и защиты практических проектов (индивидуальной или групповой), отчёта и устной защиты индивидуального вклада каждого из студентов. При этом студент может в качестве отчётной формы использовать деятельность в рамках проектной деятельности, предусмотренной образовательной программой, в рамках НУГ и научно-учебных лабораторий, индивидуальных проектов, с согласия преподавателя.

Итоговая оценка по дисциплине выставляется по следующей формуле:

$$O_{\text{итоговая}} = 0.4 \cdot (\frac{\sum O_{\text{накопленная по модулю}}}{2}) + 0.6 \cdot O_{\text{Экзамен 4 мод.}}$$

Округление арифметическое.

На любом из экзаменов студент может получить дополнительное задание или вопрос для частичной компенсации оценок текущего контроля, однако получение положительной оценки за курс без участия в проекте невозможно.

Содержание дисциплины

Раздел 1. Введение в Data Science

Темы

1. **Data Science vs Computational Social Science.** Обсуждение подходов и статей в семинарском формате, совместная выработка определения
2. **Инженерные основы Data Science** Информационные технологии. Форматы данных и DSL. Языки разметки. Языки программирования. Технологические стеки. Машинное обучение.

Статьи к разделу

- Loukides, M., 2010. What is data science? - O'Reilly Radar [WWW Document]. URL http://radar.oreilly.com/2010/06/what-is-data-science.html (accessed 12.5.14).
- Golder S.A. Digital Footprints: Opportunities and Challenges for Online Social Research / S.A. Golder, M.W. Macy // Annual Review of Sociology. – 2014. – Vol. 40. – Digital Footprints. – № 1.
- Yasseri T. Can electoral popularity be predicted using socially generated big data? / T. Yasseri, J. Bright // it - Information Technology. – 2014. – Vol. 56. – № 5. – P. 246–253.
- Social science. Computational social science / D. Lazer [et al.] // Science (New York, N.Y.). – 2009. – Vol. 323. – № 5915. – P. 721 - 723.
- Strohmaier M. Computational social science for the Word Wide Web / M. Strohmaier, C. Wagner // IEEE Intelligent Systems. – 2014. – Vol. 29. – № 5. – P. 84.
- Adamic L.A. The political blogosphere and the 2004 US election: divided they blog / L.A. Adamic, N. Glance // Proceedings of the 3rd international workshop on Link discovery. – ACM, 2005. – The political blogosphere and the 2004 US election. – P. 36–43.

Раздел 2. Современная инфраструктура обработки данных. Data Workflows

Темы

1. **Типы и структуры данных.** Векторы. Списки. Матрицы. Работа с текстовыми данными.
2. **Основы алгоритмов.**[3] Управляющие структуры. Циклы. Оператор условного перехода. Элементы функционального программирования.
3. **Основы работы с данными.** Понятие о чистых данных. Обработка и трансформация данных. Визуализация данных.
4. **Работа с реальными данными.** Чтение, предобработка и очистка данных. Современная инфрастурктура работы с текстовыми данными. Bad Data.
5. **Средства обработки и анализа данных.** Современные средства фильтрации и агрегации данных. Пакет для обработки данных 'dplyr'.
6. **Технология программирования для Data Science.** Data Science Pipeline. Организация информационных технологий для обработки данных. Организация Data Workflow. Интерфейсы, форматы и протоколы.

Литература по разделу

1. Кормен Т. и др. Алгоритмы. Построение и анализ:[пер. с англ.]. – Издательский дом Вильямс, 2009.
2. Janssens J. Data Science at the Command Line: Facing the Future with Time-Tested Tools. Data Science at the Command Line / J. Janssens. – O'Reilly Media, 2014. – 212 p.
3. McCallum, Q.E., 2012. Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work, 1 edition. ed. O'Reilly Media, Beijing; Sebastopol, CA.
4. Matloff, N.S., 2011. The art of R programming tour of statistical software design. No Starch Press, San Francisco.

[3] **Понятия и определения этого раздела вводятся по мере возникновения необходимости при изучении тем, связанных с деревьями классификации, выделением коммьюнити в соцсетевом анализе, агрегацией данных**

5. Howard, J., n.d. Designing great data products - O'Reilly Radar [WWW Document]. URL http://radar.oreilly.com/2012/03/drive
approach-.. (accessed 8.18.14).
6. O'Neil, C., Schutt, R., 2013. Doing Data Science: Straight Talk from the Frontline, 1 edition. ed. O'Reilly Media.

Раздел 3. Введение в интеллектуальный анализ данных, сетей и текстов

Темы

1. **Введение в Data Mining** Обучение с учителем. Задача классификации. Деревья решений. Оценка качества классификации. Кроссвалидация. Обучение без учителя. Задача кластеризации. Алгоритм k-средних. Иерархическая кластеризация.
2. **Введение в анализ текста.** Введение в инструментарий автоматического анализа текста. Эксплораторный анализ текстовых коллекций. Анализ корпуса на уровне документов.
3. **Основы соцсетевого анализа.** Основные понятия сетевого анализа. Меры центральности. Модели формирования и эволюции сетей. Алгоритмы Community Detection.

Литература по разделу

1. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R (Springer Texts in Statistics). 2013 edition. Springer, 2013.
2. Tan P.-N. Introduction to data mining / P.-N. Tan, M. Steinbach, V. Kumar. – Boston, Mass; London: Addison-Wesley; [Pearson Education [distributor], 2013.
3. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications / G. Miner [et al.]. – Academic Press, 2012. – 1000 p.
4. Easley D., Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world. – Cambridge University Press, 2010.
5. Kolaczyk E.D.. Statistical Analysis of Network Data with R / E.D. Kolaczyk, G. Csárdi. – New York: Springer, 2014. – 207 p.

Раздел 4. Проектная организация цифровых исследований

- Введение в проектную организацию цифровых исследований. Стадии проекта.
- Выполнение учебного проекта.
- Презентация результатов.

Оценочные средства для текущего контроля и аттестации студента

Примеры текущих заданий

Сквозное задание "Организация процесса анализа данных" Это задание предназначено для систематизации навыков организации процесса анализа данных с использованием системы data workflow Drake[4].

[4] https://github.com/Factual/drake

Литература к заданию: главы 4 и 6 из книги ”Data Science at the Command Line”, интернет-источники.

В этом задании вам необходимо организовать процесс анализа и создания отчёта по данным игровых чатов. Исходные данные извлекаются из БД в соответствии с распределением по вариантам.

Результат созданного workflow – набор отчётов в формате HTML, сгенерированных из написанных вами шаблонов RMarkdown, и содержащий:

1. общий отчёт (‘report.html‘), который должен содержать, как минимум, таблицу с общим количеством сообщений для каждого пользователя и график, показывающий помесячное количество сообщений для каждого из top-5 пользователей в вашем варианте. Приветствуется дополнительная информация, позволяющая составить общее представление об интенсивности коммуникации пользователей. Необходимо помнить, что отчёт будет создаваться динамически, поэтому привязка к конкретным данным недопустима
2. индивидуальный отчёт (‘USERID.html‘) для каждого пользователя, содержащий, как минимум, облако слов для этого пользователя, первые и последние (хронологически) пять сообщений. Не забывайте про стоп-слова. Приветствуется дополнительная информация.

Исходные данные расположены в СУБД и достаточно велики, чтобы вызвать проблемы в при непосредственной обработке в R (или сделать её очень медленной). Поэтому необходимо создание workflow, объединяющего разные технологии на разных этапах обработки, в частности вам нужно:

- Создать запрос к БД, возвращающий данные для вашего варианта. Для его генерации с учётом конкретных идентификаторов вашего варианта подготовлен скрипт на Python, который нужно добавить в Drakefile с необходимой модификацией.
- Выполнить его с использованием консольного клиента БД в скрипте оболочки
- Составить цепочку команд, извлекающую Top-20 самых частых слов в сообщениях вашего варианта (игнорируя все остальные поля, извлеченные из БД), записать Top-20 для исходного и лемматизированного с помощью ‘mystem‘ текста в файл ‘top20.txt‘ (не перезапишите файл при добавлении лемматизированных результатов!)
- Составить шаблоны RMarkdown для обоих отчётов, с использованием дополнительных источников разобраться с добавлением в ‘Drakefile‘ функции формирования отчётов по шаблону

На практическом занятии мы будем проверять работу workflow на альтернативном наборе данных. Необходимо, чтобы создание отчётов на альтернативном наборе не требовало правки workflow.

Каковы ограничения отчётов, формируемых автоматически? Как автоматизированная часть workflow встраивается в цепочку работы аналитика данных?

Пример задания по теме ”Агрегация данных” Опишите своими словами, что делают следующие функции. Напишите, какие параметры они принимают и какие результаты возвращают. Не забывайте пользоваться справкой в случае сомнений.

filter, arrange, select, summarise, group_by, mutate, n, top_n

В качестве небольшой разминки выполните следующие задания.

Дано несколько вариантов команд (фильтрация, агрегация), реализованные в “традиционном” стиле. Перепишите их, используя пайпы. Ответы впишите в пустые исполняемые блоки.

summarise(group_by(gdt, iyear), attacks=n(), killed=sum(nkill))

select(head(arrange(filter(gdt, imonth>=6, imonth<9), -nkill), 5), city, iyear, imonth, nkill)

Опишите, что показывают полученные данные (содержательно):

select(top_n(filter(filter(gdt, iyear $>=$2010), nkill $<$ nwound), 10, nwound), city, iyear, imonth, gname)

```
## Source: local data frame [11 x 4]
##
##                 city iyear imonth
## 1            Zahedan  2010      7
## 2            Hat Yai  2012      3
## 3           Damascus  2012      5
## 4  Ismail Khil district  2012      5
## 5              Sanaa  2012      5
## 6      Marmul district  2013      1
## 7           Damascus  2013      2
## 8            Beirut  2013      8
## 9            Nairobi  2013      9
## 10           Tripoli  2013     11
## 11             Sanaa  2013     12
## Variables not shown: gname (chr)
```

Получите данные в нужной форме

1. В каких странах больше всего совершается терактов? Выведите топ 5 стран и количество терактов (2 столбца).
2. (*) Выведите количество терактов в Австралии и Океании (см. столбик region_txt) по годам, упорядоченные по убыванию. Подсказка: в итоге нужно получить таблицу с двумя столбцами: год и количество терактов.
3. (*) Нужно посчитать, сколько в среднем совершается терактов за месяц (по каждому месяцу в году). Подсказка: нужно получить таблицу из двух колонок: месяц и среднее число терактов. Задайтесь сначала вопросом: ”что такое среднее число терактов в январе, феврале …?”

Пример теста по теме ”Соцсетевой анализ”

Часть I

1.

Приведите содержательный пример и опишите свойства (если по вашему мнению такой тип графа не встречается, обоснуйте):

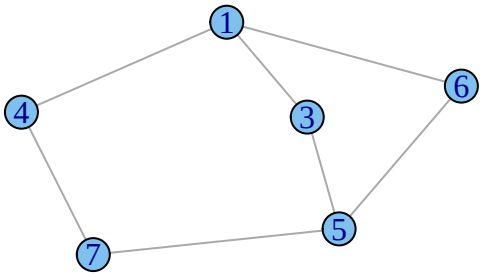
Ненаправленного графа

Направленного графа, в котором in-degree значительно отличается от out-degree

Графа с одним узлом с высокой betwenness-центральностью

Графа, содержащего узел с высокой betwenness-центральностью и degree = 2

2.



2.1. Запишите в виде матрицы смежности, edgelist и adjacency (node) list.

2.2. Каковы degree каждой из вершин? Постройте гистограмму degree.

2.3 По какой модели (Random graph (Erdos-Renyi), Preferential Attachment, Small World, …) был образован граф, по Вашему мнению? Какие свойства помогают это определить?

2.4. Для каждой из вершин 6, 4:

- выпишите проходящие через неё кратчайшие пути
- посчитайте для вершины betweenness-centrality.

3

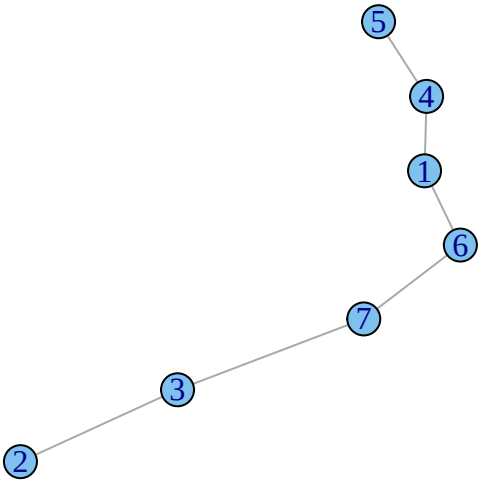
3.1. Опишите известные вам подходы к выделению community в графах

3.2. Опишите шаги алгоритма выделения коммьюнити на основе edge-betwenness

Часть II

Случайные графы

```
library(igraph)
graph <- erdos.renyi.game(7, 0.3, type=c("gnp", "gnm"),
        directed = F, loops = F)
plot(graph)
```



Вероятность узла иметь степень k задана формулой:

$$B(N-1;k;p) = \left(\frac{N-1}{k}\right)p^k(1-p)^{N-1-k}$$

$$\left(\frac{N-1}{k}\right) = \frac{(N-1)!}{k!(N-1-k)!}$$

1.1 Какого типа этот граф?

1.2 Для графа из N = 10 вершин рассчитайте вероятности узла иметь степень 5 при вероятности образования связей $p = 0.25$ и 0.75. Функция для подсчёта факториала в R: factorial(5)

1.3 Как взаимосвязаны эти вероятности для вашего графа?

1.4 Какой вид будет иметь гистограмма degree для этого графа при вероятности образования связей $p = 0.5$?

1.5 При какой средней степени вершины графа мы можем ожидать образования giant component?

2

Известно, что длина average shortest path $L_{av} \sim \frac{log(N)}{log(z)}$, где N – число узлов сети, z – средняя степень узла.

2.1 Дайте определение понятию average shortest path

2.2 При увеличении количества узлов в графе с 10^{4} до 10^{9}, L_{av} увеличится с 13 до …?

3

Приведите содержательный пример и опишите свойства:

Графа с двумя гигантскими компонентами (giant components)

Случайного графа (графа Эрдоша-Реньи)

Примеры тем проектов

- Соцсетевой анализ сообщества фанфикрайтеров
- Взаимодействие российских реп-исполнителей на основе данных о треках
- Твиттер-коммьюнити исследований в области CSS
- Соцсетевой анализ музыкальных предпочтений студентов и школьников

Другие примеры доступны по адресу http://courses.nosoc.io/current-projects.html.

Примеры проектных заданий доступны по ссылкам:

- http://courses.nosoc.io/www-assets/docs/music_proposal.pdf
- http://courses.nosoc.io/www-assets/docs/fb_proposal.pdf
- http://courses.nosoc.io/www-assets/docs/css_proposal.pdf

Методическая новизна курса и образовательные технологии

Для обеспечения необходимого уровня уверенного владения первокурсниками инструментальными средствами (языком R и средой RStudio) предусмотрен сквозной компьютерный практикум по всем разделам курса, включая элементы соцсетевого анализа. Кроме того, рабочая среда с веб-доступом позволяет прозрачно переносить работу между практикумом и самостоятельной работой студента.

Преподавание языковых средств R и концепций научного программирования осуществляется с использованием современных абстракций, с упором на понимание на концептуальном уровне и формальным введением по мере необходимости. Разделы курса, таким образом, играют двойную роль, так работа с агрегацией данных (пакеты dplyr, tidyr, magrittr) служит для практики в декомпозиции операций и неформальном введении понятий алгоритма и функции, элементы Data Mining влекут дискуссию о деревьях как структуре данных, элементы анализа соцсетей ведут к обсуждению понятий класса, объекта, свойств и методов, а также способствуют введению понятия жадного алгоритма (Гирвана-Ньюмана).

Сложность тематики для студентов-первокурсников социогуманитарных направлений и большой разброс в темпах «привыкания» студентов к работе с данными требуют новых форм и методик работы:

- широкого использования частично написанного кода (по аналогии с курсом Machine Learning Andrew Ng), что даёт студенту возможность варьировать уровень своего погружения в конкретную тему;
- использования интерактивных тьюториалов на базе swirl для дополнительной практики;
- использования технология co-teaching в сочетании с flipped classroom: предварительно записанная одним преподавателем лекция просматривается студентами заранее и обсуждается на занятии под руководством co-teacher или вместе со студентами-старшекурсниками, что создаёт практику «совместного» разбора рассматриваемой темы, и, по опыту, поощряет вопросы, которые студенты избегают задавать в традиционном лекционном формате;
- разбора методов, остающихся за рамками практической части курса в рамках семинаров по кейсам реальных исследований (как опубликованных, так и работ старшекурсников, разбираемых с авторами);
- проведения старшекурсниками гостевых тьюториалов по интересным темам, например работе с картами и првостранственными данными в R.

Объединяющим методическим компонентом курса является учебный проект, выполняемый группами студентов с применением спектра рассматриваемых в курсе методов. Проект не предполагает жесткого дисциплинарного разделения по ролям. Интересы студентов разных направлений учитываются при выборе тем (например, для студентов-экономистов и менеджеров интерес представляет анализ сети технологических компаний и их основателей на основе данных Crunchbase, одна из тем 2014-2015 года – История торговли и военных конфликтов интересна политологам и историкам, при этом большая часть тем интересна студентам разных направлений). Выполняемый под руководством одного или нескольких старшекурсников, проект задаёт минимальную технологическую планку, делая акцент на поиске дополнительных источников информации, осмысленном выборе методов, обсуждении и интерпретации результатов. При этом технически сложные элементы выполняются первокурсниками при помощи старшекурсников и/или преподавателей.

Даже при индивидуальном проекте или использовании внешнего проекта (в рамках НУЛ или НУГ, другой деятельности студента) обсуждение технических приёмов и методов, найденных дополнительных источников данных, обмен информацией со студентами, работающими по смежным темам и peer review составляют существенную часть работы.

Учебно-методическое и информационное обеспечение дисциплины

Базовый учебник

Хотя единого учебника, покрывающего все разделы дисциплины, не существует, в качестве основного пособия при выполнении практических заданий на протяжении курса используется:

1. Кабаков Р.. R в действии. Анализ и визуализация данных на языке R / Р. Кабаков. – М.: ДМК Пресс, 2013. – 580 p. (Kabacoff R.. R in action data analysis and graphics with R / R. Kabacoff. – Shelter Island, NY; London: Manning ; Pearson Education [distributor], 2011.)

Дополнительная литература

1. Chambers, John M. Software for Data Analysis Programming with R. New York; London: Springer, 2008.
2. Conway, Drew, and John Myles White. Machine Learning for Hackers. 1 edition. O'Reilly Media, 2012.
3. Downey, Allen B. Think Bayes. 1 edition. O'Reilly Media, 2013.
4. Downey, Allen B. Think Complexity: Complexity Science and Computational Modeling. 1 edition. O'Reilly Media, 2012.
5. Downey, Allen B. Think Stats. 1 edition. O'Reilly Media, 2011.
6. Gandrud, Christopher. Reproducible Research with R and RStudio (Chapman & Hall/CRC The R Series). Chapman and Hall/CRC, 2013.
7. O'Neil, Cathy, and Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. 1 edition. O'Reilly Media, 2013.
8. Wickham, Hadley. Ggplot2 Elegant Graphics for Data Analysis. Dordrecht; New York: Springer, 2009.

Программные средства и дистанционная поддержка дисциплины

Курс использует для обеспечения взаимодействия выделенный сервер на базе Linux с программными пакетами RStudio Server (Pro), Shiny Server (Pro), siwrl и другим необходимым программным обеспечением, позволяющим работу из любого места с помощью веб-браузера.

Это позволяет решить несколько задач, в том числе задачу мобильности выполнения практических заданий, обеспечения единства установленного ПО и снижения затрат времени на решение сопутствующих проблем, организацию циркуляции исходных данных и учебных материалов, и приема самостоятельных работ студентов. В дальнейшем возможна интеграция автоматической проверки работ.

В рабочей среде RStudio сервер студенты могут осуществлять все виды текущих вычислительных заданий по курсу, подготовку динамических презентаций по итогам работы и создание интерактивных приложений, визуализирующих модели.

Студентам предоставляется консольный доступ к серверу.

Прием работ, содержащих вычислительные задания, организуется путём сохранения требуемых артефактов (динамический документ, динамическая презентация, интерактивное веб-приложение, статический документ в оговоренных форматах, сопутствующие файлы данных) в специально организованный каталог рабочей среды с автоматической фиксацией времени сдачи. При этом несоблюдение требований к формату, имени и расположению артефактов, а также срока сдачи по умолчанию влечёт нулевую оценку за работу. В исключительных случаях преподаватель может принять работу, несоответствующую требованиям, или сданную после срока с пенализацией оценки по ней. Допускается приём работ по электронной почте или другим оговоренным каналам с разрешения преподавателя (в том числе если работа содержит преимущественно невычислительные задания) и в случае технических неполадок.

Исходные данные, указания к практическим и самостоятельным занятиям, тексты для домашнего чтения, как правило, размещаются в рабочей среде.

Оценки, отзывы на работы, как правило, доводятся по электронной почте или другим оговоренным каналам, включая учебные группы в соцсетях.

Используются технологии дистанционного проведения лекций с использованием Skype, Fuze Pro или аналогов (2014–2015).

Материально-техническое обеспечение дисциплины

Аудитория с проектором для лекций, компьютерные аудитории с современными версиями браузеров согласно требованиям RStudio Server к клиентам. Клиент ssh для консольного доступа к серверу.

Сервер для работы студентов (спецификация в зависимости от количества записавшихся на факультатив) с Ubuntu Linux, Shiny Server, RStudio Server и другими пакетами, в зависимости от специфики проектов.