Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет социологии
Департамент социальных наук

Рабочая программа дисциплины
Автоматизированный сбор больших данных в экономико-социологических исследованиях

для направления 39.04.01 Социология подготовки магистра
для магистерской программы
«Прикладные методы социального анализа рынков»

Разработчик программы
Управителев Ф.А., upravitelev@gmail.com

Одобрена на заседании департамента социологии «___»_____ 20 г.

Руководитель департамента А.Ю. Чепуренко _____

Рекомендована Академическим советом образовательной программы магистратуры «Прикладные методы социального анализа рынков» «___»_____ 20 г.,

№ протокола_____

Утверждена «___»_____ 20 г.

Академический руководитель образовательной программы магистратуры «Прикладные методы социального анализа рынков» Д.Х. Ибрагимова _____

Москва, 2018

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления 39.04.01 «Социология» подготовки магистра, изучающих дисциплину «Автоматизированный сбор больших данных в экономико-социологических исследованиях». Программа разработана в соответствии с:

- Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный Исследовательский Университет «Высшая Школа Экономики» по направлению 39.04.01 «Социология» подготовки магистра (http://www.hse.ru/data/2015/05/08/1098813788/ОС_маг_Социология_зам.pdf)
- Образовательной программой 39.04.01 «Прикладные методы социального анализа рынков» подготовки магистра
- Рабочим учебным планом магистерской программы «Прикладные методы социального анализа рынков»

2 Цели освоения дисциплины

Целями освоения дисциплины «Автоматизированный сбор больших данных в экономико-социологических исследованиях» является получение представления о роли данных в современном мире и формирование базовых навыков работы с большими данными – импорт данных из разных источников, чистка и манипуляции с данными.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ФГОС	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способен предлагать модели, изобретать и апробировать способы и инструменты профессиональной деятельности (формируется частично)	СК-2	выбор стратегии решения задачи	выполнение домашних заданий
Способен к самостоятельному освоению новых методов исследования, изменению научного и научно-производственного профиля своей деятельности (формируется частично)	СК-3	освоение и использование методов презентации результатов, подключений к базам данных и импорт файлов нового формата	лекции, практические занятия, выполнение домашних заданий

Способен составлять и представлять проекты научно-исследовательских и аналитических разработок в соответствии с нормативными документами (формируется частично)	ПК-4	корректное следование требованиям к оформлению кода и описанию хода решения домашних заданий	выполнение домашних заданий
Способен оформлять и представлять результаты деятельности с использованием методов, методик и приемов презентации результатов (формируется частично)	ПК-9	создание информативных и лаконичные визуальные материалы по собранным данным	выполнение домашних заданий

4 Место дисциплины в структуре образовательной программы

Для магистерской программы настоящая дисциплина является дисциплиной по выбору (Вариативная часть).
Изучение данной дисциплины базируется на следующих дисциплинах:

- Методология и методы социологического исследования
- Анализ социологических данных
- Теория вероятностей и математическая статистика

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

- Научно-исследовательский семинар
- Методология и методы исследований в социологии
- курсы по выбору

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные занятия	из них		самостоятельная работа
				Лекции	практические занятия	
1	Введение в большие данные - идеи, технологии, методы и области применения.	32	10	2	8	22
2	Виды источников данных	24	10	2	8	14
3	Методы сбора удаленных данных. Парсинг.	32	12	2	10	20
4	Методы сбора удаленных данных. Удаленные базы данных и API.	36	16	2	14	20
5	Представление результатов исследования	28	12	2	10	16
	Итого	152	60	10	50	92

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Параметры
		1	2	3	4	
Промежуточный	Домашнее задание			*		Манипуляции с данными - выделение подвыборок, слияние и объединение таблиц, создание и модификация колонок. Форма отчетности: скрипт на R
	Домашнее задание			*		Импорт предоставленных данных из csv, tsv, xlsx, sav-файлов. Форма отчетности: markdown-скрипт.
	Домашнее задание				*	Парсинг сайта, сбор данных через API. Форма отчетности: markdown-скрипт.
	Домашнее задание				*	Создание статичных и интерактивных визуализаций, презентация отчета в markdown-скрипте (*.Rmd).

7 Критерии оценки знаний, навыков

Практическая работа

По итогам самостоятельной практической работы, выполняемой на занятии в компьютерном классе, студенты должны продемонстрировать:

- способность решать поставленные задачи по импорту, анализу и визуализации данных в R
- способность представлять результаты в markdown-формате

Домашние задания

Студенты должны выполнить четыре домашних задания, после блоков по основам синтаксиса, блока по импорту данных и блока по визуализации и презентации результатов анализов. Домашние задания выполняются индивидуально, на предоставленных преподавателям материалах. При оценке домашних заданий учитывается, выполнено ли задание, читабельность и наличие комментариев в коде, общая логика решения. Скрипты домашнего задания должны выполняться без дополнительного редактирования в ОС Windows или Linux.

Каждое домашнее задание оценивается по 10-ти балльной шкале. Студенты также могут выполнить задание повышенной сложности и получить 1-3 баллов дополнительно. Оценка за курс выставляется по накопленным оценкам за выполнение практических домашних заданий.

8 Содержание дисциплины

Лекция 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.

Практикум 1.

Основы языка R. История и развитие языка, основная сфера применения. Введение в R. Установка, рабочие панели RStudio. Основы синтаксиса: операторы, вызов функций, структура выражений. Правила оформления кода.

Литература:

- An Introduction to R https://cran.r-project.org/doc/manuals/r-release/R-intro.html
- R Language Definition https://cran.r-project.org/doc/manuals/r-release/R-lang.html
- Hadley Wickham, Advanced R (part Foundations) http://adv-r.had.co.nz/
- Peng Roger D. R Programming for Data Science, part History and Overview of R https://bookdown.org/rdpeng/rprogdatascience/history-and-overview-of-r.html

Практикум 2.

Базовые типы и структуры данных: векторы, факторы, списки, таблицы. Работа с датами. Форматы дат, unixtimestamp и стандарт ISO8601.

Литература:

- Kabacoff, R. (2015). R in Action: Data Analysis and Graphics with R. http://kek.ksu.ru/eos/DataMining/1379968983.pdf

Практикум 3.

Условные операторы - if...else, ifelse, switch. Циклы и векторизованные функции. Собственные функции.

Литература:

- Шипунов, А. Б., Балдин, Е. М., Волкова, П. А., Коробейников, А. И., Назарова, С. А., Петров, С. В., & Суфиянов, В. Г. (2012). Наглядная статистика. Используем R!. *М.: ДМК Пресс*, *298*, 1.https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf

Практикум 4.

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.

Литература:

- Руководство по data.table https://bookdown.org/statist_/DataTableManual/
- Data.table Reference semantics https://cran.r-project.org/web/packages/data.table/vignettes/datatable-reference-semantics.html
- A data.table R tutorial by DataCamp: intro to DT[i, j, by] https://www.datacamp.com/community/tutorials/data-table-r-tutorial#gs.xKK3HNU
- Advanced tips and tricks with data.table http://brooksandrew.github.io/simpleblog/articles/advanced-data-table/
- Cheat Sheet Data.table https://s3.amazonaws.com/assets.datacamp.com/blog_assets/datatable_Cheat_Sheet_R.pdf

Практикум 5.

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

Литература:

- Data Guidelines https://f1000research.com/for-authors/data-guidelines
- R Markdown http://rmarkdown.rstudio.com/

Лекция 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.

Практикум 6.

Импорт текстовых (txt, csv) файлов. Ошибки при импорте. Сохранение данных.

Литература:

- R Data Import/Export https://cran.r-project.org/doc/manuals/r-release/R-data.html
- This R Data Import Tutorial Is Everything You Need part I https://www.datacamp.com/community/tutorials/r-data-import-tutorial/#gs.WdNbNT0
- Data Import Cheat Sheet https://github.com/rstudio/cheatsheets/raw/master/source/pdfs/data-import-cheatsheet.pdf

Практикум 7.

Импорт специализированных форматов – xls, xlsx (пакет readxl) и sav (пакет foreign). Сохранение меток при импорте файлов SPSS (sav).

Литература:

- This R Data Import Tutorial Is Everything You Need part II https://www.datacamp.com/community/tutorials/importing-data-r-part-two#gs.EODdys8

Лекция 3. Методы сбора удаленных данных. Парсинг.

Сайты как источник данных. HTML, XPath, DOM-разметка. CSS-селекторы. Пакет rvest.

Практикум 8.

Простейшие парсеры. Парсинг таблиц и разделов.

Литература:

- Пакет rvest: easy web scraping with R https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/
- Краткое руководство по XPath http://soltau.ru/index.php/themes/dev/item/413-kratkoe-rukovodstvo-po-xpath
- Web Scraping and Parsing Data in R https://www.datacamp.com/community/tutorials/exploring-h-1b-data-with-r#gs.qBLAbWo

Лекция 4. Методы сбора удаленных данных. Удаленные базы данных и API.

Подключение и импорт данных из базы данных. Облачная архитектура. Подключение к API. Oauth-авторизация. Шифрование данных. Пароли. Хранение персональных данных, законы о защите персональных данных.

Практикум 9.

Доступ к удаленным базам данных. Основные коннекторы. Примеры SQL-запросов.

Литература:

- Тренировочные задания и учебники по SQL http://www.sql-ex.ru/

- Справочник функций и команд SQL https://www.w3schools.com/sql/

Практикум 10.

Импорт данных через API. Чтение документации, подключение через токены. Подключение к API сервисов погоды.

Литература:

- HTTP request methods https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods
- API Tutorial for Beginners https://blog.cloudrail.com/api-tutorial-for-beginners/

Практикум 11.

Импорт данных социальных сетей. Подключение к Facebook API

Литература:

- Package "Rfacebook" https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf
- Facebook Graph API documentation https://developers.facebook.com/docs/graph-api/
- Analyze Facebook with R http://thinktostart.com/analyzing-facebook-with-r/

Практикум 12.

Импорт данных социальных сетей. Подключение к Vkontakte API

Литература:

- VK API documentation https://vk.com/dev/manuals
- Package 'vkR' https://cran.r-project.org/web/packages/vkR/vkR.pdf

Лекция 5. Визуализация данных

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.

Литература:

- Principles of Information Display for Visualization Practitioners http://www2.cs.uregina.ca/~rbm/cs100/notes/spreadsheets/tufte_paper.html
- Data looks better naked http://www.darkhorseanalytics.com/blog/data-looks-better-naked
- Clear Off the Table http://www.darkhorseanalytics.com/blog/clear-off-the-table/
- Подбор правильных цветовых палитр для визуализации данных https://infogra.ru/infographics/podbor-pravilnyh-tsvetovyh-palitr-dlya-vizualizatsii-dannyh

Практикум 13

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

Литература:

- Hadley Wickham, A Layered Grammar of Graphics http://byrneslab.net/classes/biol607/readings/wickham_layered-grammar.pdf
- ggplot2 cheatsheet https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

Практикум 14

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.

Литература:

- Data Driven Documents (d3.js documentation) https://github.com/d3/d3/wiki
- The plotly cookbook https://plotly-book.cpsievert.me/the-plotly-cookbook.html

9 Оценочные средства для текущего контроля и аттестации студента

Примеры домашних заданий

1. Импортируйте *.csv-файл, с учетом нестандартных разделителей
2. Импортируйте *.sav-файл с сохранением меток
3. Импортируйте таблицу данных из удаленной PostgreSQL-базы
4. Визуализируйте динамику цен на недвижимость с линией сглаженного среднего
5. Визуализируйте интерактивную тепловую карту количества покупок в онлайн-магазине по часам и по районам Москвы

10 Порядок формирования оценок по дисциплине

Итоговая оценка по дисциплине складывается из накопленных оценок за домашние задания. Экзамен не проводится. В ходе курса студенты могут получить 40 баллов за четыре домашних задания (по 10 за каждое). В том случае, если студент выполнял задания повышенной сложности, суммарная накопленная оценка может быть больше 40 баллов.

Для получения оценки по 10-ти балльной шкале сумма набранных баллов делится на 4 и округляется арифметически. Если из-за выполненных заданий повышенной сложности оценка оказалась выше 10 баллов, то выставляется оценка в 10 баллов.

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, последнее домашнее задание принимается не позднее, чем за неделю до начала сессии четвертого модуля. Оценки за курс выставляются в течение сессии четвертого модуля.

11 Учебно-методическое и информационное обеспечение дисциплины

11.1 Литература и интернет-ресурсы

- A data.table R tutorial by DataCamp: intro to DT[i, j, by]
https://www.datacamp.com/community/tutorials/data-table-r-tutorial#gs.xKK3HNU
- Advanced tips and tricks with data.table
http://brooksandrew.github.io/simpleblog/articles/advanced-data-table/
- An Introduction to R https://cran.r-project.org/doc/manuals/r-release/R-intro.html
- Analyze Facebook with R http://thinktostart.com/analyzing-facebook-with-r/
- API Tutorial for Beginners https://blog.cloudrail.com/api-tutorial-for-beginners/
- Cheat Sheet Data.table
https://s3.amazonaws.com/assets.datacamp.com/blog_assets/datatable_Cheat_Sheet_R.pdf
- Clear Off the Table http://www.darkhorseanalytics.com/blog/clear-off-the-table/
- Data Driven Documents (d3.js documentation) https://github.com/d3/d3/wiki
- Data Guidelines https://f1000research.com/for-authors/data-guidelines
- Data Import Cheat Sheet
https://github.com/rstudio/cheatsheets/raw/master/source/pdfs/data-import-cheatsheet.pdf
- Data looks better naked http://www.darkhorseanalytics.com/blog/data-looks-better-naked

- Data.table Reference semantics https://cran.r-project.org/web/packages/data.table/vignettes/datatable-reference-semantics.html
- Facebook Graph API documentation https://developers.facebook.com/docs/graph-api/
- ggplot2 cheatsheet https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- Hadley Wickham, A Layered Grammar of Graphics http://byrneslab.net/classes/biol607/readings/wickham_layered-grammar.pdf
- Hadley Wickham, Advanced R (part Foundations) http://adv-r.had.co.nz/
- HTTP request methods https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods
- Package “Rfacebook” https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf
- Package “vkR” https://cran.r-project.org/web/packages/vkR/vkR.pdf
- Peng Roger D. R Programming for Data Science, part History and Overview of R https://bookdown.org/rdpeng/rprogdatascience/history-and-overview-of-r.html
- Principles of Information Display for Visualization Practitioners http://www2.cs.uregina.ca/~rbm/cs100/notes/spreadsheets/tufte_paper.html
- R Data Import/Export https://cran.r-project.org/doc/manuals/r-release/R-data.html
- R Language Definition https://cran.r-project.org/doc/manuals/r-release/R-lang.html
- R Markdown http://rmarkdown.rstudio.com/
- The plotly cookbook https://plotly-book.cpsievert.me/the-plotly-cookbook.html
- This R Data Import Tutorial Is Everything You Need part I https://www.datacamp.com/community/tutorials/r-data-import-tutorial/#gs.WdNbNT0
- This R Data Import Tutorial Is Everything You Need part II https://www.datacamp.com/community/tutorials/importing-data-r-part-two#gs.EODdys8
- VK API documentation https://vk.com/dev/manuals
- Web Scraping and Parsing Data in R https://www.datacamp.com/community/tutorials/exploring-h-1b-data-with-r#gs.qBLAbWo
- Краткое руководство по XPath http://soltau.ru/index.php/themes/dev/item/413-kratkoe-rukovodstvo-po-xpath
- Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com С.97-123
- Пакет rvest: easy web scraping with R https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/
- Подбор правильных цветовых палитр для визуализации данных https://infogra.ru/infographics/podbor-pravilnyh-tsvetovyh-palitr-dlya-vizualizatsii-dannyh
- Руководство по data.table https://bookdown.org/statist_/DataTableManual/
- Справочник функций и команд SQL https://www.w3schools.com/sql/
- Тренировочные задания и учебники по SQL http://www.sql-ex.ru/

11.2 Дистанционная поддержка дисциплины

Все материалы (презентации лекций, материалы к практическим занятиям, тексты статей или ссылки на онлайн-материалы) высылаются студентам на адрес групповой электронной почты и в slack-канал.

11.3 Материально-техническое обеспечение дисциплины

В ходе аудиторных занятий используется ноутбук и проектор для демонстрации слайдов. Практические занятия проходят в компьютерном классе. Студенты обеспечиваются необходимыми файлами или доступами к базам данных для работы на практических занятиях и подготовки домашних заданий.