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

Факультет компьютерных наук
Департамент больших данных и информационного поиска

Рабочая программа дисциплины «Введение в культуру данных»

для образовательной программы «Программа двух дипломов НИУ ВШЭ и Лондонского университета «Международные отношения»
направления подготовки 41.03.05 «Международные отношения»
уровень бакалавр

Разработчик(и) программы
Деркач Д. А. PhD, dderkach@hse.ru
Кантонистова Е. О, к. ф.-м. н. ekantonistova@hse.ru
Казеев Н. А. nkazeev@hse.ru

Одобрена на заседании департамента больших данных и информационного поиска
«___»_____ 201_ г.
Руководитель департамента
В. В. Подольский_____ [подпись]

Утверждена Академическим советом образовательной программы
«___»_____ 201_ г., № протокола_____

Академический руководитель образовательной программы s
М. В. Братерский _____ [подпись]

Москва, 2017

Undergraduate Program in International Relations
Introduction to Data Science

Course objectives:

This course provides an introduction to modern data analysis methods through a series of lectures and practice work based on Excel. The aims of the course are to:

- provide an introduction to modern data science techniques;
- introduce main concepts of scientific data analysis;
- show best practices of working with data:
- train basic skills in MS Excel.

Learning outcomes:

At the end of the course, and having completed the set readings and the activities, students will be able to:

- demonstrate knowledge of basic concepts of data science;
- formulate and solve simple scientific analysis problems;
- perform a data analysis in MS Excel.

Course language: English

Course description:

This course offers an introduction to the modern Data Science methods that are useful both for research and industrial career. The main focus of the course is to give a full breadth of topics that can be attributed to data science without going into details. Instead students are trained to develop critical thinking and scientific approach to problem solving.

The course starts from the discussion of data science applications in science with some examples explained. This covers both analytical and ethical questions of the solutions. A part of the course also concerns the main methods of data storage and their usage.

The second part concerns the main methods that lead to scientific results of the analyses in humanities starting from simple time series analyses to the simplest applications of the machine learning techniques. The lectures are followed by the practical tutorials in MS Excel.

Section 1: Introduction to Data Science.

Weeks 1-2:

Applied data science in the international relations, examples of applications, examples of erroneous applications. Data. Big data. Distributions, basic types of distributions. Main characteristics of the distributions: mean, median, moving values. Correlations and causality.

Weeks 3-4:

Main methods of accumulation, storage and processing of data. Types of variables. Prospects for the development and application of data analysis. Ethical aspects of the development of artificial intelligence. Processing polls, ratings. The simplest text analyses.

Section 2: Basic Data Analysis.

Weeks 5-6:

Time series. Forecasting. Basic methods for forecasting irregular time series. Accounting for periodic factors.

Weeks 7-8:

Event selection, outlier search, selection cross-checks, methods for improving selection, observable engineering. Optimization. Fisher discriminant, linear models of discrimination.

Weeks 9-10:

Data analysis. Hypothesis, quality metrics, blind analysis. Hypothesis testing. p-value. Confidence interval.

Section 3: Introduction to Machine Learning.

Weeks 11-12:

Machine learning as a tool for searching for regularities. The main tasks of Machine Learning: regression, classification, clustering, visualization. Quality metrics. Data types. Terminology: object, target variable, attribute, quality metric, model, training method. Examples of statements of problems in the humanities. Analysis of specific problems, features, quality metrics.

Weeks 13-14:

Methods of machine learning. Overview of the main types of models and the principles of their learning (on simple examples). Decision trees and forests. Neural networks.

Weeks 15-16:

Artificial intelligence and neural networks. Neural networks as a sequence of data transformations. The idea of learning neural networks. Architecture of neural networks. The main applications of neural networks.

Methods and Forms

The methods and forms of study used on this course include lectures, practice work, teachers' consultations and homework. There is one biweekly lecture and one weekly seminar. Each seminar is based on the contents of the preceding lecture. Seminar includes practical work in computer class under supervision of teacher. Homework normally includes the computer-based assignment.

Duration: modules 3 and 4, Spring 2018.

Course structure: A biweekly lecture followed by a weekly seminar.

Final assessment. Homework, Classwork, Examination.

Module Grade:

Final assessment = 0.8*Ongoing Assesment + 0.2*Examination Assessment,
Where
Ongoing Assesment = 0.3*Tests + 0.5*Homework + 0.2*Midterm Examination

For more information please check the Russian version of the program below.

***Instructors*:**
Dr. Denis Derkach (dderkach@hse.ru)
Dr. Elena Kantonistova (ekantonistova@hse.ru)
Mr. Nikita Kazeev (nkazeev@hse.ru).

Recommended Literature:

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer series in statistics.
- Bishop C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Support Material:

- MS Office Blog https://blogs.office.com/en-us/?eu=true
- Support documentation for Microsoft Excel: https://support.office.com/en-US/Excel
- R Studio http://www.r-studio.com/

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

Факультет компьютерных наук
Департамент больших данных и информационного поиска

Рабочая программа дисциплины «Введение в культуру данных»

для образовательной программы «Программа двух дипломов НИУ ВШЭ и Лондонского университета «Международные отношения»
направления подготовки 41.03.05 «Международные отношения»
уровень бакалавр

Разработчик(и) программы
Деркач Д. А. PhD, dderkach@hse.ru
Кантонистова Е. О, к. ф.-м. н. ekantonistova@hse.ru
Казеев Н. А. nkazeev@hse.ru

Одобрена на заседании департамента больших данных и информационного поиска
«___»_____ 201_ г.
Руководитель департамента
В. В. Подольский_____ [подпись]

Утверждена Академическим советом образовательной программы
«___»_____ 201_ г., № протокола_____

Академический руководитель образовательной программы s
М. В. Братерский _____ [подпись]

Москва, 2017

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 дисциплину «Введение в культуру данных», учебных ассистентов и студентов направления подготовки 41.03.05 «Международные отношения», обучающихся по образовательной программе «Программа двух дипломов НИУ ВШЭ и Лондонского университета «Международные отношения».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ по направлению подготовки 41.03.05 Международные отношения;

2 Цели освоения дисциплины

Целями освоения дисциплины «Введение в культуру данных» являются:

- ознакомление студентов с основами науки о данных;
- формирование у студентов практических навыков работы с данными, решения прикладных задач анализа данных по специальности и визуализации данных;
- формирование у студентов навыков поиска информации.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

Уровни формирования компетенций:

РБ — ресурсная база, в основном теоретические и предметные основы (знания, умения);

СД – способы деятельности, составляющие практическое ядро данной компетенции;

МЦ – мотивационно-ценностная составляющая, отражает степень осознания ценности компетенции человеком и готовность ее использовать

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной	УК-1	РБ	Усваивает новую информацию, воспроизводит информацию с дополнительными вопросами, применяет выученные методы на практике, обобщает полученную информацию, экстраполирует полученные знания на нестандартные примеры.	выполнение заданий на семинарах, выполнение домашней работы	Проверка домашней работы, экзамен
Способен работать с информацией: находить, оценивать и использовать ин-	УК-5	РБ, СД	Получать информацию из определённого источника, сравнивать несколько источников информации, искать несоответствия меж-	Выполнение домашней работы,	Проверка домашней работы, оценивание самостоя-

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
формацию из различных источников, необходимую для решения научных и профессиональных задач (в том числе на основе системного подхода)			ду источниками информации, самостоятельно искать источники информации, оценивать качество источников информации	самостоятельная работа на семинаре	тельной работы на семинаре
Способен критически оценивать и переосмыслять накопленный опыт (собственный и чужой), рефлексировать профессиональную и социальную деятельность	УК-9	МЦ	Искать закономерности, обнаружать ошибки исследования, корректировать анализ	Прослушивание лекций, выполнение домашней работы, самостоятельная работа на семинаре	Проверка домашней работы, экзамен
Способен проводить анализ информации, эффективно используя современные технологии сбора и хранения информации	ПК-2	РБ, СД	Искать информацию, сохранять и систематизировать информацию, разрабатывать способ автоматизации анализа и хранения информации	Прослушивание лекций, выполнение домашней работы, самостоятельная работа на семинаре	Проверка домашней работы, экзамен
Способен давать интерпретацию, а также корректно применять результаты анализа международных проблемах в профессиональной деятельности	ПК-4	МЦ	Интерпретировать данные, проверять совместимость теории и данных, проверять гипотезы, интерпретировать чужие исследования	Прослушивание лекций, выполнение домашней работы, самостоятельная	Проверка домашней работы, экзамен

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
				работа на семинаре	
Способен самостоятельно собирать и обрабатывать информацию из различных источников по конкретной, определенной руководителем научной тематике в сфере международных отношений	ПК-20	РБ, СД	Извлекать данные из доступных хранилищ, готовить данные к анализу, искать новые источники данных, сравнивать разные источники данных	Прослушивание лекций, выполнение домашней работы, самостоятельная работа на семинаре	Проверка домашней работы, экзамен
Способен научно интерпретировать данные отечественной и зарубежной статистики о социально-экономических и политических процессах и явлениях	ПК-22	РБ, СД	Интерпретировать данные, оценивать разные подходы к сбору и анализу данных	Прослушивание лекций, выполнение домашней работы, самостоятельная работа на семинаре	Проверка домашней работы, экзамен

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к блоку дисциплин общего цикла.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Высшая математика и статистика

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

- базовая компьютерная грамотность;
- базовое знание математики.

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

- анализ данных в R;

- выполнение КР и ВКР.

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы				Самостоятельная работа
			Лекции	Семинары	Практические занятия	Другие виды работы[1]	
1	Введение в науку о данных	29	4	8	-	-	17
2	Простейшие алгоритмы анализа данных	43	6	12	-	-	25
3	Введение в машинное обучение	42	6	12	-	-	24
	ИТОГО	114	14	32	0	0	66

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры **
		3	4	
Текущий	Домашнее задание	1	1	Работа на компьютере по закреплению пройденного материала
Промежуточный	Коллоквиум	*		Устный
Итоговый	Экзамен		*	Письменный

7 Критерии оценки знаний, навыков

В курсе предусмотрено несколько форм контроля знания:

- Самостоятельные работы на семинарах, проверяющие знание основных фактов с лекций.
- Практические домашние работы Excel, формирующие навыки работы с основными инструментами анализа данных, а также помогающие освоить основные концепции машинного обучения
- Устный коллоквиум в конце 1-го модуля
- Письменный экзамен.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

[1] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

8 Содержание дисциплины

Раздел представляется в удобной форме (список, таблица). Изложение строится по разделам и темам. Содержание темы может распределяться по лекционным и практическим занятиям.

Раздел 1 Введение в науку о данных

Лекция 1: Наука о данных в применении к международным отношениям, примеры применения, примеры ошибочного применения. Данные. Большие данные. Распределения, основные типы распределений. Основные характеристики распределений: среднее, медиана, скользящие величины. Корреляции и причинность.

Лекция 2: Основные методы накопления, хранения и обработки данных. Типы переменных. Перспективы развития и применения анализа данных. Этические аспекты развития искусственного интеллекта. Обработка опросов, рейтинги. Простейшие тексты.

Практикум 1: Знакомство с Excel, основные функции. Контроль МООС курса (не оценивается). Построение таблиц, чтение документов в разных форматах

Практикум 2: Вычисление базовых характеристик распределения. Среднее, медиана. Корреляции.

Практикум 3: Получение данных из сети Интернет. Обработка опросов.

Литература:

- https://support.office.com/en-us/article/AVERAGE-function-047bac88-d466-426c-a32b-8f33eb960cf6
- https://support.office.com/en-us/article/CORREL-function-995dcef7-0c0a-4bed-a3fb-239d7b68ca92
- МООС курсы:
  - https://www.youtube.com/watch?v=eI_7oc-E3h0
  - https://www.youtube.com/watch?v=pPSwbK4_GdY
  - https://www.youtube.com/watch?v=RgvdCHjOKYg

Практикум 4: Обработка текстов, вычисление рейтингов.

Литература:

- https://support.microsoft.com/en-us/help/214153/how-to-count-the-occurrences-of-a-number-or-text-in-a-range-in-excel
- https://support.office.com/en-us/article/RANK-function-6a2fc49d-1831-4a03-9d8c-c279cf99f723

1. Раздел 2 Простейшие алгоритмы анализов данных

Лекция 3: Временные ряды. Прогнозирование. Основные методы прогноза нерегулярных временных рядов. Учёт периодических факторов.

Лекция 4: Отбор событий, выбор аномалий, проверка отбора, методы улучшения отбора, инжиниринг наблюдаемых. Оптимизация. Дискриминант Фишера, линейные модели дискриминации.

Лекция 5: Анализ данных. Гипотеза, метрика качества решения, слепой анализ. Подтверждение гипотез. p-value. Вычисление p-value. Доверительный интервал.

Практикум 5: Вычисление скользящих величин. Анализ временных рядов в Excel. Нахождение тренда, сезонности.

Литература:

- https://www.youtube.com/watch?time_continue=1&v=gHdYEZA50KE
- https://support.office.com/en-us/article/Analyze-trends-in-data-using-sparklines-be6579cf-a8e3-471a-a459-873614413ce1

Практикум 6: Анализ временных рядов (продолжение). Предсказание поведения временного ряда в будущем.

Литература:

- http://www.excel-easy.com/examples/moving-average.html
- https://www.youtube.com/watch?v=ynXkxPnosJQ

Практикум 7: Отбор событий, выбор аномалий, проверка отбора, методы улучшения отбора, инжиниринг наблюдаемых.

Литература:

- https://support.office.com/en-us/article/Create-a-box-and-whisker-chart-62f4219f-db4b-4754-aca8-4743f6190f0d
- https://support.office.com/en-us/article/Find-and-remove-duplicates-00E35BEA-B46A-4D5D-B28E-66A552DC138D

Практикум 8: Линейный анализ данных в Excel.

Литература:

- *https://support.office.com/en-ie/article/Data-Analysis-7E71735C-C471-47E1-84EF-A8C23DC3098B*

Практикум 9: Оптимизация. Дискриминант Фишера, линейные модели дискриминации.

Литература:

- http://www.real-statistics.com/multivariate-statistics/discriminant-analysis/

2. Раздел 3 Введение в машинное обучение

Лекция 6: Машинное обучение как инструмент автоматического поиска закономерностей. Основные постановки задач: регрессия, классификация, кластеризация, визуализация. Обучение на прецедентах и обучающая выборка. Метрики качества. Типы данных. Терминология: объект, целевая переменная, признак, метрика качества, модель, метод обучения. Примеры постановок задач в гуманитарных науках. Разбор конкретных постановок, признаков, метрик качества на этих задачах.

Лекция 7: Методы машинного обучения. Обзор основных типов моделей и принципов их обучения (на простых примерах). Решающие деревья, решающие леса. Нейронные сети.

Лекция 8: Искусственный интеллект и нейронные сети. Нейронные сети как последовательность преобразований данных. Идея обучения нейронных сетей. Архитектура нейронных сетей. Основные применения нейронных сетей.

Практикум 10: Оценка качества готового решения. Кластеризация.

Литература:

- Andrew Ng https://www.coursera.org/learn/machine-learning
- https://en.wikipedia.org/wiki/Receiver_operating_characteristic
- https://ru.coursera.org/learn/business-analytics-decision-making/lecture/pGFT5/4-cluster-analysis-with-excel

Практикум 11: Регрессия. Визуализация результатов.
Литература:
http://www.excel-easy.com/examples/regression.html

Практикум 12: Задача классификации. Построение решающего дерева.
Литература:
https://help.xlstat.com/customer/en/portal/articles/2062258-classification-tree-in-excel-tutorial?b_id=9283

Практикум 13: Нейронные сети. Примеры использования, построение предсказаний.

- https://www.wordseye.com/workspace
- https://experiments.withgoogle.com/ai/teachable-machine

Практикум 14: Нейронные сети. Построение простейших нейронных сетей.
Литература:

- https://www.youtube.com/watch?v=3993kRqejHc

9 Образовательные технологии

В первой части курса используется Microsoft Excel, доступ к которому обеспечен студентам НИУ ВШЭ через портал office365. Остальные программные пакеты доступны для бесплатного скачивания из сети Интернет.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Домашние и контрольные работы выполняются во внеаудиторное время с использованием программы Excel.

10.2 Примеры заданий промежуточной аттестации

Примеры вопросов для домашней работы:

- Подсчитайте среднее, медиану для набора наблюдений.
- Для набора данных, полученных за последний месяц, постройте предсказание на следующую неделю.
- Постройте линейную регрессию.

10.3 Примеры вопросов итоговой аттестации

- Основные характеристики распределений: среднее, медиана, скользящие величины. Корреляции и причинность.
- Основные постановки задач машинного обучения. Примеры прикладных задач.
- Гипотеза, метрика качества, оценка справедливости гипотезы.

11 Порядок формирования оценок по дисциплине

Результирующая оценка по дисциплине рассчитывается по формуле:

О_итог=0.8 О_накопл + 0.2 О_экз

Накопленная и итоговая оценки округляются арифметически.

Накопленная оценка рассчитывается по формуле:

О_накопл=0.3 О_самост + 0.5 О_дз + 0.2 О_коллоквиум

Оценка за самостоятельную работу вычисляется как сумма баллов по всем самостоятельным, переведенная в 10 бальную шкалу. Оценка за домашнюю работу — как сумма баллов по всем практическим заданиям и соревнованию, переведенная в 10 бальную шкалу. Количество баллов за разные задания может различаться в зависимости от их сложности. Все промежуточные оценки (за домашние, самостоятельные и коллоквиум) могут быть не целыми. Накопленная и итоговая оценки округляются математически.

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Основная литература

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer series in statistics.
- Bishop C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

12.2 Дополнительная литература и ресурсы

MS Office Blog https://blogs.office.com/en-us/?eu=true
MS Office Support https://support.office.com/en-us

12.3 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства: Microsoft Excel.

12.4 Дистанционная поддержка дисциплины

Все материалы (слайды лекций, практических занятий, дополнительные инструкцию) доступны в сети Интернет по ссылке, высылаемой студентам.

13 Материально-техническое обеспечение дисциплины

В ходе аудиторных занятий используется персональный компьютер и проектор для демонстрации слайдов. Практические занятия проходят в компьютерном классе с доступом в сеть

интернет. Студенты обеспечиваются необходимыми файлами или инструкциями для их получения для работы на практических занятиях и подготовки домашних заданий.