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

Факультет Санкт-Петербургская школа экономики и менеджмента
Национального исследовательского университета
«Высшая школа экономики»

Департамент прикладной математики и бизнес-информатики

Рабочая программа дисциплины
Научно-исследовательский семинар "Вероятностные методы моделирования"

для образовательной программы «Анализ больших данных в бизнесе, экономике и обществе»
направления подготовки 01.04.02«Прикладная математика и информатика»
уровень магистратура

Разработчик(и) программы
Сироткин А.В., к.ф.-м.н., доцент, avsirotkin@hse.ru

Согласована менеджером ОП Анализ больших данных в бизнесе, экономике и обществе

Е.С. Авдониной _____

«30» августа 2016г.

Утверждена Академическим руководителем образовательной программы

А.В. Сироткин _____

«30» августа 2016г.

Санкт-Петербург, 2016

1 Область применения и нормативные ссылки

Настоящая рабочая программа дисциплины устанавливает минимальные требования к знаниям и умениям студента, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину, учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе».

Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ http://www.hse.ru/data/2016/11/02/1111123560/01.04.02%20Прикладная%20математика%20и%20информатика.pdf;
- Образовательной программой «Анализ больших данных в бизнесе, экономике и обществе», направление подготовки 01.04.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программ «Анализ больших данных в бизнесе, экономике и обществе».

2 Цели освоения дисциплины

Целями освоения дисциплины Научно-исследовательский семинар "Вероятностные методы моделирования" являются:

- научить студентов азам научно-исследовательской деятельности;
- научить студентов структурировать исследование и взаимосвязывать различные разделы исследования;
- научить студентов выделять самое основное для презентации результатов исследования и оформлять презентационные материалы.

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 следующие 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
Инструментальные компетенции					
Способен организовывать научно-исследовательскую деятельность.	ПК-9	РБ, СД	Может составить план научной работы и придерживаться его.	Семинарские занятия, самостоятельная работа студентов	Выступление, экзамен
Способен создавать междисциплинарные тексты с ис-	ПК-11	СД, МЦ	Умеет готовить научные доклады и научные отчеты.	Семинарские занятия, самостоятельная работа	Выступление, эссе, 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
пользованием языка и аппарата прикладной математики.				студентов	
Способен публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий).	ПК-12	СД	Способен презентовать результаты научной деятельности, как на основе своей работы, так и на основе научных статей, написанных другими авторами.	Семинарские занятия, самостоятельная работа студентов	Выступление, аудиторная работа, экзамен
Способен в составе научно-исследовательского и производственного коллектива решать задачи профессиональной деятельности в соответствии с профилем подготовки, общаться с экспертами в других предметных областях.	ПК-19	РБ,СД	Умеет уточнять постановку задачи анализа данных в диалоге с заказчиком или коллегами поставившими первичную задачу. Соблюдает сроки выполнения экспериментов и предоставления отчетных материалов.	Семинарские занятия, самостоятельная работа студентов	Выступление, экзамен

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин проектной и исследовательской работы и блоку дисциплин, обеспечивающих магистерскую подготовку.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении любых дисциплин, связанных с научно-исследовательской деятельностью студентов.

5 Тематический план учебной дисциплины

ОБЪЕМ ДИСЦИПЛИНЫ - 9 зачетные единицы

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Научно-исследовательская работа студентов – виды, содержание, особенности	12		4		8
2	Выбор направления и формулировка темы исследования. Постановка целей и задач. Гипотезы.	12		4		8

	Предмет и объект исследования.					
3	Работа с источниками, цитирование, оформление ссылок и списка литературы	12		4		8
4	Эмпирические/полевые/иные исследования - сбор материалов для практической части работы	12		4		8
5	Методы и модели – особенности, выбор, использование, совмещение	12		4		8
6	Структура работы, логика и взаимосвязь, использование иллюстративного материала, оформление	12		4		8
7	Представление итогов - речь, презентация, раздаточные материалы, правила выступления.	12		4		8
8	Введение в проблемы дата майнинга	12		4		8
9	Алгоритмы сетевого анализа.	28		8		20
10	Восстановление скрытых распределений пользователей в Вконтакте	16		8		8
11	Медиа – войны в интернете	12		4		8
12	Межстрановые исследования.	12		4		8
13	Выделение патернов поведения из больших данных.	18		8		10
14	Исследование Инстаграм	12		4		8
15	Анализ профилей пользователя и выявление скрытых особенностей	14		6		8
16	Применение классификаторов для предсказания котировок	12		4		8
17	Цели и задачи классификации медицинских данных	12		4		8
18	Обзор методов Sentiment analysis	20		8		12
19	Требования и структура исследовательского проекта.	12		4		8

20	Методы исследования. Содержание и логика научной работы.	12		4		8
21	Обсуждение будущей письменной работы и её защиты.	14		6		8
22	Защита и презентация научной работы.	12		4		8
23	Подготовка, защита, презентация научной работы	12		4		8
24	Обсуждение статей	28		8		20
		342		120		222

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Параметры
		1	2	3	4	
Текущий	Выступление		8			Представление темы исследования, в форме обзорного доклада.
	Аудиторная работа		*	*	*	Доклады по отдельным статьям на темы согласованные с преподавателем
	Эссе			*		Письменная работа
Итоговый	Экзамен				*	Экзамен в форме публичного доклада о результатах проведенного исследования.

7 Критерии оценки знаний, навыков

В процессе освоения курса предусмотрены следующие формы контроля:

- текущий контроль на семинарах: выступления с докладами на темы согласованные с преподавателем.

- текущий контроль во втором модуле: выступление с докладом о теме исследования, эссе, описывающее текущее состояния области исследования, выбранной студентом.

- итоговый контроль: экзамен, в форме публичного доклада о результатах научных исследований полученных в течении первого года обучения на программе «Анализ больших данных в бизнесе, экономике и обществе»

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

Текущий контроль в форме выступления:

При оценивании берутся во внимание:

- самостоятельность выполнения
- соблюдение основных правил подготовки и представления материалов
- грамотная речь
- умение отвечать на вопросы

Эссе.

НИС ориентирован на формирование у студентов навыков проведения самостоятельных исследований, поэтому дисциплина строится на семинарских занятиях. Основной упор делается на самостоятельную работу студентов. В ходе прохождения дисциплины студенты должны выполнить индивидуально.

Электронные файлы с текстами эссе проверяются на оригинальность (объемы заимствований из других текстов) через систему antiplagiat.ru. При обнаружении системой заимствований более 20% текста эссе не оценивается.

При выставлении оценки за реферат учитывается:

- понимание проблематики в рамках выбранной темы;
- знание контекста, материала;
- степень самостоятельности студента в оценивании исследуемой проблемы, независимости от чужого мнения;
- оригинальность рассуждений;
- умение анализировать чужую точку зрения и средства ее выражения;
- умение аргументировано излагать свою точку зрения;
- умение выстроить свой текст (композиция, логика);
- обоснованность даваемых в работе выводов и рекомендаций (если таковые имеются);
- богатство и точность языка;
- грамотность;
- единство стиля.

Таким образом, студенты смогут отработать следующие навыки: применение профессиональных знаний и умений; ведение исследовательской работы; реализация критического мышления; публичное выступление. Кроме проверки освоенности компетенций, студенты тренируются правильно оформлять свои научные работы.

8 Содержание дисциплины

1. Научно-исследовательская работа студентов – виды, содержание, особенности.
2. Выбор направления и формулировка темы исследования. Постановка целей и задач. Гипотезы. Предмет и объект исследования.
3. Работа с источниками, цитирование, оформление ссылок и списка литературы.
4. Эмпирические/полевые/иные исследования - сбор материалов для практической части работы.
5. Методы и модели – особенности, выбор, использование, совмещение.
6. Структура работы, логика и взаимосвязь, использование иллюстративного материала, оформление.
7. Представление итогов - речь, презентация, раздаточные материалы, правила выступления.
8. Введение в проблемы дата майнинга:

10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH. QIANG YANG, Department of Computer Science Hong Kong University of Science and Technology Clearwater Bay, XINDONG WU Department of Computer Science University of Vermont

9. Алгоритмы сетевого анализа. Обсуждение различных моделей, понятие модулярити.

Community detection in graphs, Santo Fortunato, Complex Networks and Systems Lagrange Laboratory.

10. Восстановление скрытых распределений пользователей в Вконтакте. Для чего это нужно и как это можно сделать? Обсуждение проблемы шума в сырых данных. Задание на сбор данных и анализ полученных результатов. Работа с датасетом из Вконтакта.

Demographic research with non-representative internet data, Emilio Zagheni Department of

Sociology, University of Washington, Seattle, Washington, USA, and Ingmar Weber Department of Social Computing, Qatar Computing Research Institute, Doha, Qatar
The Privacy Jungle: On the Market for Data Protection in Social Networks Joseph Bonneau, Computer Laboratory, University of Cambridge Sören Preibusch
11. Медиа – войны в интернете, на примере сравнения контента российских и украинских каналов. Задание на сбор данных, Анализ проблем препроцессинга русского языка, проведения тематического моделирования, анализ полученных результатов. Обучение работе с программой TopicMiner. Работа с Российским и украинским датасетом.
12. Межстрановые исследования. Обсуждение проблем препроцессинга китайского языка. Cross-Cultural Analysis of Blogs and Forums of UK, India, Singapur.
Задание на обработку китайского датасета.
Assessing Censorship on Microblogs in China. King-wa Fu, Chung-hong Chan, and Michael Chau. 2013
Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models, Michael Paul and Roxana Girju, University of Illinois at Urbana Champaign.
Задание на анализ наиболее известных токенайзеров и попытка проведения процедуры токенизации на основе китайского датасета.
A. Stanford word segmenter http://nlp.stanford.edu/software/segmenter.shtml
B. ICTCLAS http://repos.6estates.com/nexus/content/groups/public/com/nus/ictclas-tool/
Huaping Zhang, Hongkui Yu, Deyi Xiong, Qun Liu. 2003. HHMM -based Chinese C. **Lexical Analyzer ICTCLAS**. In Proceedings of 2nd SIGHAN Workshop on Chinese Language Processing, pp.184-187
FNLP (Fudan NLP tool by Xipeng Qiu) http://jkx.fudan.edu.cn/~xpqiu/
13. Выделение патернов поведения из больших данных.
Catch Me If You Can: Detecting Pickpocket Suspects from Large-Scale Transit Records.
Bowen Du State Key Lab of Software Development Environment Beihang University
14. Исследование Инстаграм. Обсуждение датасета и генерирование идей исследования. Проведение исследование. Данные доступны по результатам летней школы Digital Traces (https://eu.spb.ru/digitaltraces2016/main)
What We Instagram: A First Analysis of Instagram Photo Content and User Types
Yuheng Hu Lydia Manikonda Subbarao Kambhampati Department of Computer Science, Arizona State University
Visualizing Instagram: Tracing Cultural Visual Rhythms. Nadav Hochman History of Art and Architecture University of Pittsburgh,
Raz Schwartz Human Computer Interaction Institute, Carnegie Mellon University
15. Анализ профилей пользователя и выявление скрытых особенностей. Анализ профилей на основе работы Ingmar Weber. Возможность применения данных из школы Digital Traces (https://eu.spb.ru/digitaltraces2016/main). Возможность репликации работы Вебера на основе данных из Вконтакта.
Crowdsourcing Health Labels: Inferring Body Weight from Profile Pictures. Ingmar Weber, Qatar Computing Research Institute, Yelena Mejova Qatar Computing Research Institute.
Social Media Image Analysis for Public Health. Venkata Rama Kiran Garimella Aalto University Helsinki, Finland.
16. Применение классификаторов для предсказания котировок акций Газпрома. Где и как достать данные, препроцессинг. Краткий обзор классификаторов для анализа котировок. Задание по предсказанию котировок.
17. Цели и задачи классификации медицинских данных. Обсуждение датасета отзывов по врачам. Задание по классификации отзывов.
Scope of Data Mining in Medicine, Divdeep Singh Sukhpreet Kaur, M.Tech CSE Assistant Professor Department of Computer Science and Engineering Department of Computer Science and Engineering Sri Guru Granth Sahib World University Sri Guru Granth Sahib World University

Uniqueness of medical data mining, Krzysztof J. Ciosa,b,c,d, G. William Mooree,f,g a Department of Computer Science and Engineering, University of Colorado at Denve
What Affects Patient (Dis)satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Mode, Michael J. Paul Dept. of Computer Science Johns Hopkins University.

18. Обзор методов Sentiment analysis. Реализация классификаторов на основе данных из проекта РГНФ (http://linis-crowd.org/). Задание по сентимент анализу.

Sentiment Strength Detection in Short Informal Text, Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton
Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media, S. Alexeeva, S. Kolcov, O. Koltsova National Research Institute Higher School of Economics.

19. Требования и структура исследовательского проекта.

Цели, задачи, методы проведения исследования. Требования к научным работам. Основные принципы исследовательской деятельности. Разбор действующих документов в НИУ ВШЭ - Санкт-Петербург на предмет оформления работы. Приведение примеров для лучшего усвоения материала.

20. Методы исследования. Содержание и логика научной работы.

Характеристика основных структурных элементов. Рассмотрение постановки научной цели (или целей), а также вытекающих из нее (из них) важных задач. Логика научной работы - специфика и необходимость. Взаимосвязь информационной базы и применяемых методов исследования, выявление специфики исследовательской базы на различных рынках. Раскрытие взаимосвязи тематики исследования и используемых для этого методов.

21. Обсуждение будущей письменной работы и её защиты.

Основные характеристики письменной научной работы (эссе/реферат). Обсуждение возможных сложностей и ошибок. Выявление возможных сильных и слабых сторон будущей работы. Выбор правильных ориентиров для сбора и обработки информации. Обсуждение и проработка вопросов обработки недостоверной информации.

При обсуждении тем будущих работ особое внимание уделяется способности каждого студента находить информацию для выбранной темы исследования, а также умению аргументировано отстаивать свою точку зрения.

22. Защита и презентация научной работы.

Научная работа может быть защищена с использованием презентаций. В презентации должны быть представлены научные результаты. Желательным элементом является дискуссия. Важным моментом является критическое восприятие и умение корректно заимствовать сильные стороны работ сокурсников.

23. Подготовка, защита, презентация научной работы

9 Образовательные технологии

Проводится представление докладов по тематике предложенной преподавателем или студентом по теме, согласованной с семинаристом.

В рамках семинара проводятся выступления приглашенных специалистов из разных областей.

9.1 Методические рекомендации преподавателю

Для НИСа предпочтительно использовать новые публикации в ведущих международных и российских рецензируемых журналах. Студенты выполняют эссе по статьям. Перечень статей составляется преподавателем.

В эссе, приводится обзор текущего состояния области исследования, выбранной студентом.

9.2 Методические указания студентам по освоению дисциплины

1. Организация работы над эссе:

1.1. В течение 3-го модуля 2016/17 уч. года предполагается написание эссе.

1.2. Цели выполнения эссе:

– овладение методами поиска, анализа, переработки и систематизации информации по заданной теме;

– развитие умения осмыслить и изложить точку зрения других авторов, и на их основе сформулировать свои выводы.

1.3. Срок сдачи эссе устанавливается преподавателями.

Эссе, сданные позднее, не оцениваются.

1.4. В любой момент до срока сдачи эссе можно сдать или прислать по электронной почте для предварительной консультации черновик эссе.

1.5. Этапы написания эссе:

1.5.1. Согласование темы эссе с преподавателем.

1.5.2. Поиск литературных источники, в которых отражающих текущее состояние исследований в данной области. Можно использовать как «бумажные» источники, так и Интернет-публикации. Составить список литературы.

1.5.3. Ознакомиться с точкой зрения различных ученых (прочитать выбранные работы), при необходимости выписать цитаты, зафиксировав их источник (полное описание книги или статьи, номер страницы, на которой приведена цитата).

1.5.4. Набросать черновой вариант работы.

1.5.5. Одобрить предварительный вариант у преподавателя.

1.5.6. Написать окончательный вариант работы. Сдать не позднее назначенного срока.

2. Какие источники использовать для написания эссе.

2.1. Источники, которые следует использовать при написании эссе, – книги (монографии), учебники, статьи в научных журналах, аналитические и справочные интернет-ресурсы на русском и английском языке.

2.2. Форма источников – на твердом носителе (книги, журналы и т.п.) и ресурсы из сети Интернет (тексты статей, аналитических обзоров и т.п.).

При оценке эссе будут применяться следующие критерии:

1. Подбор и систематизация материалов по теме. Отражение в работе основной проблематики по направлению темы исследования. Качество литературных источников и их соответствие теме.
2. Глубина понимания исследуемой проблемы. Оперирование ключевыми понятиями и владение терминологией. Знание фактического материала. Полнота раскрытия темы.
3. Структурированность работы, выстраивание логики изложения. Постановка цели и задач работы и качество их решения. Обоснование актуальности темы исследования и выводы по результатам работы.
4. Степень самостоятельной переработки материала источников. Умение выделять основные идеи анализируемых работ, применять критический анализ основных положений, сопоставлять разные точки зрения, строить выводы, аргументировано и связно излагать свои мысли.
5. Стиль и грамотность изложения.

9.3 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

- Web of Science, https://www.webofknowledge.com/
- Scopus, https://www.scopus.com/

10 Оценочные средства для текущего контроля и аттестации студента

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

Примеры статей, которые можно использовать при подготовки индивидуальных выступлений:

1. **Community detection in graphs**, Santo Fortunato, Complex Networks and Systems Lagrange Laboratory.
2. **Demographic research with non-representative internet data**, Emilio Zagheni Department of Sociology, University of Washington, Seattle, Washington, USA, and Ingmar Weber Department of Social Computing, Qatar Computing Research Institute, Doha, Qatar
3. **Sentiment Strength Detection in Short Informal Text**, Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton

11 Порядок формирования оценок по дисциплине

Накопленная оценка по дисциплине рассчитывается с помощью взвешенной суммы оценок за отдельные формы текущего контроля знаний следующим образом:

$$O_{накопленная} = 0{,}2 \cdot O_{эссе} + 0{,}2 \cdot O_{выст} + 0{,}6 \cdot O_{ауд} \text{ , где}$$

$O_{выст}$ – оценка за выступление с обзорным докладом по теме исследования

$O_{ауд}$ – суммарная оценка за выступления с отдельными докладами по статьям согласованным с преподавателями

$O_{эссе}$ – оценка за эссе

Результирующая оценка по дисциплине рассчитывается следующим образом:

$$O_{результ} = 0{,}6 \cdot O_{накопл} + 0{,}4 \cdot O_{экз} \text{ , где}$$

$O_{накопл}$ – накопленная оценка по дисциплине

$O_{экз}$ – оценка за экзамен

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Основная литература

Курс построен на изучении современных статей по теме исследования студентов и не имеет обязательной для всех литературы.

12.2 Дополнительная литература

1. Demographic research with non-representative internet data / Emilio Zagheni, Ingmar Weber [Electronic Resource] // International Journal of Manpower. 2015. Vol. 36 (1). p. 13-25. - Authorized access: http://www.emeraldinsight.com/doi/pdfplus/10.1108/IJM-12-2014-0261 (Online Digital Library "Emerald eJournals").
2. Bonneau, Joseph, Preibusch,Sören.The Privacy Jungle: On the Market for Data Protection in Social Networks [Electronic Resource] // Economics of Information Security and Privacy / Tyler Moore, David Pym, Christos Ioannidis. - New York : Springer, 2010. - p. 121-167. - Author-

ized access: http://link.springer.com/chapter/10.1007/978-1-4419-6967-5_8 (Online Digital Library "Springer eBook").

3. Crowdsourcing Health Labels: Inferring Body Weight from Profile Pictures / Ingmar Weber, Yelena Mejova [Electronic Resource]. - Mode of access: https://arxiv.org/pdf/1602.07185v1.pdf (Open e-print database “arXiv”)
4. Social Media Image Analysis for Public Health / Venkata Rama, Kiran Garimella, Abdulrahman Alfayad, Ingmar Weber [Electronic Resource]. – Mode of access: https://arxiv.org/pdf/1512.04476v2.pdf (Open e-print database “arXiv”)

12.3 Справочники, словари, энциклопедии

База данных зарубежной периодики: www.jstor.org - издания по экономике, бизнесу, социологии, статистике, математике.

12.4 Ресурсы информационно-телекоммуникационной сети «Интернет»

12.5 Программные средства

- R
- Python

12.6 Информационные справочные системы

- Web of Science, https://www.webofknowledge.com/
- Scopus, https://www.scopus.com/

12.7 Дистанционная поддержка дисциплины

В образовательной среде LMS могут быть размещены различные учебные материалы, задания, литература по курсу; также в LMS возможно сдавать эссе, домашние работы, проводить и проверять текущие и итоговые тестирования.

13 Материально-техническое обеспечение дисциплины

1. Для проведения занятий необходим мультимедиа проектор, ноутбук или стационарный компьютер, экран, выход в сеть Интернет, доступ к электронным ресурсам НИУ ВШЭ.
2. Технические средства для показа слайдов, выполненных в Power Point или в формате pdf файла.