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

Аспирантская школа по образованию

Рабочая программа дисциплины

«Большие данные и методы машинного обучения в исследованиях образования»

для образовательной программы «Образование»
по направлению подготовки научно-педагогических кадров в аспирантуре 44.06.01 Образование и педагогические науки

Разработчики программы:
Докука София Владимировна, кандидат социологических наук, sdokuka@hse.ru
Сивак Елизавета Викторовна, esivak@hse.ru
Смирнов Иван Борисович, кандидат наук об образовании, ibsmirnov@hse.ru

Утверждена Академическим советом Аспирантской школы по образованию
«11» октября 2018 г., протокол № 34

Академический директор
Е.А. Терентьев _____
(подпись)

Москва – 2018

1. Область применения и нормативные ссылки

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

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и аспирантов, обучающихся по образовательной программе «Образование» и изучающих дисциплину «Большие данные и методы машинного обучения в исследованиях образования».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ по направлению подготовки кадров высшей квалификации 44.06.01 Образование и педагогические науки».
- Образовательной программой «Образование».
- Учебным планом образовательной программы «Образование».

2. Цели и задачи освоения дисциплины

Цель освоения дисциплины «Большие данные и методы машинного обучения в исследованиях образования» состоит в формировании у аспирантов:

- знаний наиболее актуальных работ в области применения новых типов данных в социальных науках и науках об образовании;
- навыков по сбору данных из социальных медиа и других цифровых следов с использованием языка программирования Python;
- навыков обработки и анализа различных типов данных (сетевые, текстовые и геоданные) с использованием языка программирования Python.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины аспирант должен:

Знать:

- Основные теоретические, методологические и практические подходы к анализу больших данных и новых типов данных;
- Источники новых типов данных;
- Ключевые исследовательские работы и направления в области применения больших данных и методов машинного обучения в социальных науках и науках об образовании;
- Основные этические принципы работы с данными и этические проблемы, связанные с использованием больших данных;

Уметь:

- Ставить исследовательские вопросы и формулировать гипотезы, протестировать которые можно с использованием больших данных;
- Грамотно использовать алгоритмы машинного обучения и статистического анализа для изучения больших данных;
- Проводить исследование полного цикла с использованием новых типов данных;
- Интерпретировать и оформлять полученные результаты.

Иметь навыки (приобрести опыт):

- Постановки исследовательского вопроса в области применения больших данных, данных нового типа и методов машинного обучения в исследованиях образования;
- Планирование исследования полного цикла;
- Презентации и защиты индивидуального исследовательского проекта.

В результате освоения дисциплины аспирант осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способность генерировать оригинальные теоретические конструкции, гипотезы и исследовательские вопросы	УК-2	Демонстрирует способность самостоятельно формулировать и обосновывать исследовательские вопросы и гипотезы	Обсуждения в ходе семинаров, подготовка и презентация итогового проекта исследования
Способность выбирать и применять методы исследования, адекватные предмету и задачам исследования	УК-3	Демонстрирует способность выбирать и применять методы исследования, адекватные предмету и задачам исследования, при выполнении заданий к занятиям 2-7 и подготовке итогового проекта исследования	Семинарские занятия, выполнение практических заданий к занятиям 2-7, подготовка и презентация итогового проекта исследования. Чтение научных статей по результатам эмпирических исследований.
Способность собирать, анализировать, обрабатывать и хранить данные в соответствии с общепринятыми научными и этическими стандартами	УК-4	Использует язык Python, демонстрирует умение работать с разными видами данных	Семинарские занятия, самостоятельная работа с данными при выполнении заданий к занятиям 2-7
Способность самостоятельно осуществлять научно-исследовательскую деятельность в области образования с использованием современных методов исследования и информационно-коммуникационных технологий	ОПК-1	Демонстрирует навыки использования информационно-коммуникационных технологий при обработке данных.	Семинарские занятия; работа над домашними заданиями, включающими необходимость проанализировать данные. Чтение научных статей по результатам эмпирических исследований.
Способность к выполнению междисциплинарных исследований в области социально-психологических, социокультурных, социально-экономических аспектов образования	ПК-2	Демонстрирует способность к выполнению междисциплинарных исследований в области социально-психологических, социокультурных, социально-экономических аспектов образования при выполнении заданий к занятиям 2-7 и подготовке итогового проекта исследования	Семинарские занятия и работа над домашними заданиями, включающими необходимость не только проанализировать данные, но и дать содержательную интерпретацию; подготовка собственного проекта исследования. Чтение научных статей по резуль-

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
			татам эмпирических исследований.
Способность решать прикладные задачи развития образовательной организации с использованием результатов современных исследований в области образования и смежных областях	ПК-4	Демонстрирует способность решать прикладные задачи развития образовательной организации с использованием результатов эмпирических и теоретических исследований в ходе подготовки и презентации собственного проекта исследования.	Семинарские занятия, работа над собственным проектом исследования, Чтение научных статей, презентация проекта исследования.
Способность критически оценивать собственные результаты в контексте результатов современных педагогических, социально-психологических, социокультурных, социально-экономических исследований	ПК-6	Демонстрирует способность критически оценивать результаты собственных исследований в контексте результатов современных исследований в области образования.	Подготовка проекта исследования, чтение научных статей.

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин по выбору и изучается на 2-м году обучения.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Дизайн и методы научного исследования в образовании
- Методология исследования и базовая статистика

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

- Способность генерировать оригинальные теоретические конструкции, гипотезы и исследовательские вопросы
- Способность выбирать и применять методы исследования, адекватные предмету и задачам исследования

5. Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	

1	Большие данные и методы машинного обучения в социальных науках и исследованиях образования. Новые типы данных: интернет-данные, другие цифровые следы и возможности их применения (лекция). Обсуждение идей индивидуальных исследовательских проектов (семинар).	9	1	2	0	6
2	Введение в язык программирования python. Базовые типы данных. Переменные. Операторы. Условия, циклы и функции. Ошибки и предупреждения.	8	1	0	1	6
3	Автоматический сбор данных из интернета. Форматы данных. HTML и JSON. Использование API интернет-сервисов на примере социальной сети ВКонтакте.	8	1	0	1	6
4	Анализ социальных сетей: основные теоретические понятия и приложения. Изучение сетей дружбы на примере данных «ВКонтакте».	8	1	0	1	6
5	Использование методов машинного обучения для предсказания характеристик пользователей на основании их цифровых следов.	8	1	0	1	6
6	Интеллектуальный анализ текстов. Основные теоретические понятия и приложения. Тематическое моделирование. Анализ текстов из социальной сети «ВКонтакте».	8	1	0	1	6
7	Анализ геопространственных данных. Основные теоретические понятия и приложения. Методы сбора, практическое использование и интерпретация результатов.	8	1	0	1	6
8	Этика использования больших данных. Алгоритмы и дискриминация.	9	1	2	0	6
9	Презентация индивидуального исследовательского проекта.	10	0	2	0	8
	Итого:	76	8	6	6	56

6. Формы контроля знаний аспирантов:

Тип контроля	Форма контроля	Параметры **
Текущий	Текущий	Активность на занятиях
Итоговый	Зачет	Презентация итогового проекта исследования

7. Критерии оценки знаний, навыков

Текущий контроль проводится в форме оценивания активности на занятиях. Активность оценивается по 10-балльной шкале. Критерии оценивания: объем и релевантность вопросов темам курса, объем участия в дискуссия, качество ответа на вопросы преподавателей (насколько ясен и обоснован ответ).

Итоговый контроль проводится в форме оценивания описания проекта исследования. Описание проекта исследования должно включать: 1) формулировку исследовательского вопроса, 2) его обоснование, 3) источник данных, описание выборки, подходы к анализу данных. Проект оценивается по 10-балльной шкале. Критерии оценивания:

0-3 – задание не выполнено или почти не выполнено

4-6 – задание сдано, но выполнено не полностью (не описаны все требуемые пункты; неясно сформулирован исследовательский вопрос; есть несоответствия между вопросом и выбранным методом или данными и др.)

7-8 – задание выполнено полностью, в т.ч. корректно описаны все требуемые пункты,

9-10 – заметны приложенные дополнительные усилия аспиранта (использована дополнительная релевантная литература, представлены подходы и методы, не рассмотренные в курсе, и др.).

8. Содержание дисциплины

Тема 1. Новые типы данных: Интернет-данные, другие цифровые следы и возможности их применения.

Основная литература

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Hobbs, W. R., Burke, M., Christakis, N. A., & Fowler, J. H. (2016). Online social integration is associated with reduced mortality risk. Proceedings of the National Academy of Sciences, 113(46), 12980-12984.

Palchykov, V., Kaski, K., Kertész, J., Barabási, A. L., & Dunbar, R. I. (2012). Sex differences in intimate relationships. Scientific Reports, 2, 370.

Тема 2. Краткое введение в Python. Переменные, списки и словари. Условия, циклы, функции.

Электронные ресурсы:
https://stepik.org/course/67 Базовый курс по Python
https://stepik.org/course/512/ Более продвинутый курс по Python
https://stepik.org/course/568/ Адаптивный тренажер по программированию на Python.
https://python.swaroopch.com/ A Byte of Python: учебник по Python для начинающих.

https://gawron.sdsu.edu/python_for_ss/course_core/book_draft/index.html Jean Mark Gawron, Python for Social sciences. Учебник для начинающих.
https://realpython.com/ Учебные материалы для изучающих Python.

Тема 3. Форматы данных. HTML и JSON. Использование API ВКонтакте.

Электронные ресурсы:
https://vk.com/dev/
https://realpython.com/python-json/
https://stepik.org/course/512/ Раздел «Применение Python: анализ текста», занятия про HTML и JSON.
https://habr.com/post/280238/

Тема 4. Информация о пользователях и дружеские связи. Что такое социальная сеть? Какие метрики используются для описания как социальной сети в целом, так и ее отдельных элементов? Центральности.

Основная литература

Newman, M. E. (2003). The structure and function of complex networks. SIAM review, 45(2), 167-256.

Smirnov, I., Sivak, E., & Kozmina, Y. (2016). In Search of Lost Profiles. Educational Studies Moscow No 4, 2016.

Smirnov, I., & Thurner, S. (2017). Formation of homophily in academic performance: Students change their friends rather than performance. PloS one, 12(8), e0183473.

Дополнительная литература

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university press.

Электронные ресурсы
http://networksciencebook.com/

Тема 5. Интересы учащихся. Анализ подписок на группы.

Основная литература

Поливанова К. Н., Смирнов И. Б. Что в профиле тебе моем: Данные «ВКонтакте» как инструмент изучения интересов современных подростков // Вопросы образования. 2017. № 2. С. 134-152.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. Social networks, 30(4), 330-342.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences, 109(1), 68-72.

Тема 6. Работа с текстом. Анализ публичных записей пользователей.

Основная литература

Smirnov I. The Digital Flynn Effect: Complexity of Posts on Social Media Increases over Time, in: Social Informatics 9th International Conference, SocInfo 2017 Proceedings. Springer, 2017. P. 24-30

Kern, M.L., Eichstaedt, J.C., Schwartz, H.A., Park, G., Ungar, L.H., Stillwell, D.J., Kosinski, M., Dziurzynski, L. and Seligman, M.E., 2014. From “Sooo excited!!!” to “So proud”: Using language to study development. Developmental psychology, 50(1), p.178.

Sivak E. V., Smirnov I. Gender Bias in Sharenting: Both Men and Women Mention Sons More Often Than Daughters on Social Media / Cornell University. Series math "arxiv.org". 2018.

Тема 7. Работа с геоданными: алгоритмы предобработки и анализа.

Основная литература

Hasan, S., Zhan, X. and Ukkusuri, S.V., 2013, August. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing (p. 6). ACM.

Centellegher, S., De Nadai, M., Caraviello, M., Leonardi, C., Vescovi, M., Ramadian, Y., ... & Lepri, B. (2016). The Mobile Territorial Lab: a multilayered and dynamic view on parents’ daily lives. EPJ Data Science, 5(1), 3.

Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. (2016) The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 4:e2537

Тема 8. Этика использования новых данных.

Основная литература

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 201320040.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. Big Data & Society, 3(1).

Henderson, M., Johnson, N. F., & Auld, G. (2013). Silences of ethical practice: dilemmas for researchers using social media. Educational research and evaluation, 19(6), 546-560.

Zimmer M. "But the data is already public": on the ethics of research in Facebook // Ethics and information technology. 2010. Vol. 12(4).

9. Оценочные средства для текущего контроля и аттестации аспиранта

Для текущего контроля оценивается активность на занятиях.

Задание для итоговой оценки качества освоения дисциплины представляет собой подготовленный проект исследования.

10. Порядок формирования оценок по дисциплине

Оценивается активность на занятиях (на всех занятиях в целом), а также проект исследования.

Итоговый контроль:

Оценивается итоговый проект исследования (текст объемом до двух страниц).

Результирующая оценка за итоговый контроль выставляется по 10-бальной шкале.

Итоговая оценка складывается из суммы текущей оценки и оценки за проект исследования и переводится в 10-бальную шкалу. В диплом выставляет результирующая оценка по учебной дисциплине, которая формируется по следующей формуле:

$$O_{дисциплина} = 0.2\ O_{текущий} + 0.8 \cdot O_{итоговый\ проект\ исследования}$$

Способ округления результирующей оценки по учебной дисциплине: арифметический (например, оценка 4,4 округляется до 4, а оценка 4,5 до 5).

11. Учебно-методическое и информационное обеспечение дисциплины

Базовый учебник

Не предусмотрен.

Основная литература

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.
Hobbs, W. R., Burke, M., Christakis, N. A., & Fowler, J. H. (2016). Online social integration is associated with reduced mortality risk. Proceedings of the National Academy of Sciences, 113(46), 12980-12984.
Palchykov, V., Kaski, K., Kertész, J., Barabási, A. L., & Dunbar, R. I. (2012). Sex differences in intimate relationships. Scientific Reports, 2, 370.
Newman, M. E. (2003). The structure and function of complex networks. SIAM review, 45(2), 167-256. https://epubs.siam.org/doi/pdf/10.1137/S003614450342480?xid=PS_smithsonian&
Smirnov, I., Sivak, E., & Kozmina, Y. (2016). In Search of Lost Profiles. Educational Studies Moscow No 4, 2016.
Smirnov, I., & Thurner, S. (2017). Formation of homophily in academic performance: Students change their friends rather than performance. PloS one, 12(8), e0183473.
Поливанова К. Н., Смирнов И. Б. Что в профиле тебе моем: Данные «ВКонтакте» как инструмент изучения интересов современных подростков // Вопросы образования. 2017. № 2. С. 134-152.
Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. Social networks, 30(4), 330-342.
Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences, 109(1), 68-72.

Программные средства

Для успешного освоения дисциплины, аспирант использует следующие программные средства:
1) Браузер Chrome (или другой современный браузер)
2) Пакет Anaconda (https://www.anaconda.com)
3) Офисные пакеты (Microsoft Office, Libre Office и т. п.) или их облачные аналоги

12. Материально-техническое обеспечение дисциплины

Ноутбук (компьютер) для преподавателя;
Проектор (для лекций или семинаров);
Ноутбук для каждого аспиранта.