Правительство Российской Федерации

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

Факультет гуманитарных наук
Школа лингвистики

Программа дисциплины «Введение в цифровые гуманитарные науки»

для студентов бакалавриата факультета гуманитарных наук

Авторы программы:
Бонч-Осмоловская А.А., кандидат филологических наук, abonch@gmail.com
Фишер, Франк, PhD, ffisher@gmail.com
Орехов, Б.В., кандидат филологических наук, borekhov@hse.ru

Рекомендована Академическим советом образовательной программы
«01» июня 2016 г., № протокола 10

Утверждена «01»июня 2016 г.
Академический руководитель образовательной программы
Ю.А. Ландер

Москва, 2016

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 цифровые гуманитарные науки» в рамках майнора «Современные методы в гуманитарных науках», учебных ассистентов и студентов факультета гуманитарных наук НИУ ВШЭ.

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ;
- Образовательной программой «Фундаментальная и компьютерная лингвистика».
- Образовательной программой «Филология».

Цели освоения дисциплины

Цель освоения дисциплины — познакомить студентов с новыми исследовательскими подходами и задачами в истории, литературоведении, лингвистике, культурологии, опирающимися на применение методов компьютерной обработки текста. Эти методы включают в себя создание и анализ корпусов текстов, построение статистических моделей, работу с большими данными, формализацию параметров текста, картографирование и т.п. Упомянутые подходы широко используются в современной гуманитарной науке, задача курса состоит в том, чтобы показать студентам, что можно с их помощью исследовать, и научить технике того, как это делать.

Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- *Знать* современные методы получения, обработки и анализа гуманитарных данных
- *Уметь* использовать в исследовательской работе корпуса текстов, свободные инструменты текстовой аналитики и картографирования
- *Иметь навыки (приобрести опыт)* корпусных исследований, количественного анализа художественных текстов, анализа географических данных и социальных сетей в гуманитарных областях

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ
Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной	УК-1
Способен работать с информацией: находить, оценивать и использовать информацию из различных источников, необходимую для решения научных и профессиональных задач (в том числе на основе системного подхода)	УК-5
Способен работать в команде	УК-7
Способен проводить сбор и документацию лингвистических данных	ПК-6

Компетенция	Код по ФГОС/ НИУ
Способен участвовать в создании представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, лингвистических и социолингвистических баз данных и пользоваться этими ресурсами	ПК-11
Способен ориентироваться в системе общечеловеческих ценностей и ценностей мировой и российской культуры, понимать значение гуманистических ценностей для сохранения и развития современной цивилизации	ПК-24

Место дисциплины в структуре образовательной программы

Настоящая дисциплина является частью майнора «Современные методы в гуманитарных науках».

Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Обзор новых методов в гуманитарных науках		2	2		
2	Компьютерные методы в филологии: «дальнее чтение» Ф. Моретти, макроанализ М. Джокерса		2	2		
3	Компьютерные методы в филологии: стилеметрия (компьютерная стилистика), тематическое моделирование		4	2	2	5
4	Компьютерные методы в лингвистике: корпусные исследования		4	2	2	
5	Теория сетей (графов) и её применения в гуманитарных науках		6	6	4	5
6	Открытые картографические пакеты и сервисы, их применение в гуманитарных науках;		4	2	2	5
7	Цифровые научные (критические) издания художественной литературы		8	4		
8	Культуромика (количественные методы исследования культуры). «Большие данные» в гуманитарных науках. Анализ данных Википедии и Twitter. Визуализация данных.		10	6	4	5
		100	40	26	14	20

Формы контроля знаний студентов

Итоговый экзамен проводится в форме защиты проекта (proof of concept) цифрового гуманитарного исследования или ресурса.

Критерии оценки знаний, навыков

В рамках курса предполагается два коллоквиума и одна итоговая презентация (защита); кроме того, студенты делают доклады на семинарах, в ходе которых демонстрируют освоение и проработку выбранного для освещения материала. Освоение части тем предполагает выполнение домашних заданий.

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

Содержание дисциплины

Тема 1. Обзор новых методов в гуманитарных науках.

Гуманитарные науки и вызов цифровой эпохи. Анализ данных в литературоведении, лингвистике, исторической науке, культурологии, истории искусства и креативных индустрий. Цифровые издания, картографические проекты, количественное отображение культурных трендов, визуализация, сети и графы в гуманитарных науках.

4 часа аудиторной работы.

Liu, 2011; Hoover, 2007; Moretti, 2005; Schreibman *et al*, 2004

Тема 2. Компьютерные методы в филологии: «дальнее чтение» Ф.Моретти, макроанализ М. Джокерса.

Проблема репрезентативности литературного канона. «Дальнее чтение» (Distant Reading) Франко Моретти как способ объективизации филологического исследования. Макроанализ М. Джокерса. Идея «масштабируемого чтения». Использование инструмента Google Ngram Viewer; ограничения этого инструмента.

4 часа аудиторной работы.

Moretti, 2013/Моретти, 2016; Jockers, 2013; Michel *et al*, 2011; Ярхо, 2006

Тема 3. Компьютерные методы в филологии: стилеметрия (компьютерная стилистика), тематическое моделирование.

История формальных исследований авторского стиля. Проблема спорного авторства и количественные подходы к её решению. Спорное авторство произведений Шекспира, «Записок федералиста», Музиля, Шолохова и д.р. Пример раскрытия авторства Дж. Роулинг (Juola, 2006). Понятие идиостиля автора. Программный пакет stylo для определения авторства (Eder *at al*, 2013).

Вероятностное тематическое моделирование. Латентное размещение Дирихле. Тематическое моделирование текстов на естественном языке на основе мешка слов. Программный пакет MALLET.

8 часов аудиторной работы.

Mendenhall, 1887; Mosteller, Wallace, 1963; Burrows, 2002; Juola, 2006; Eder *at al*, 2013; Мухин, 2011; Blei, 2012; Воронцов, 2013

Тема 4. Корпусные исследования.

Понятие корпуса. Корпус текстов как исследовательский объект. Виды корпусов. Лингвистические корпусы, национальные корпусы. Параллельные, устные, мультимедийные корпусы. Разметка и метаинформация в корпусе. Запросы к корпусу. Типология корпусных исследований.

Корпусные исследования прозы: на какую помощь со стороны цифрового знания рассчитывают литературоведы? Что можно посчитать в романе? Корпусные исследования поэзии. Семантический ореол метра.

4 часа аудиторной работы.

Hoover *et al*, 2014; Orekhov, 2015; Ляшевская, 2016

Тема 5. Теория сетей (графов) и её применения в гуманитарных науках.

Математический граф и его основные элементы (ребра, вершины). Применение теории графов в анализе социальных сетей. История становления сетевого анализа в гуманитарных областях. Гарвардский прорыв 1960-х. Социологические, исторические, культурологические иследования с использованием теории графов. Применение сетевого анализа в литературоведении.

Значимые количественные (математические) параметры графа. Плотность и диаметр графа. Основные свойства ребер и вершин. Степень вершины. Метрики центральности (betweenness centrality, closeness centrality).

Программа для анализа и визуализации графов Gephi. Основные возможности Gephi: импорт графа, алгоритмы укладки на плоскости, подсчет метрик и статистики, кластеризация графа с использованием Лувенского алгоритма.

16 часов аудиторной работы.

Newman, 2005; Bastian *et al*, 2009; Moretti, 2011;

Тема 6. Открытые картографические пакеты и сервисы, их применение в гуманитарных науках.

Основные компоненты геоинформационных систем. Географические данные в гуманитарных науках. Карты прошлого: исторические геоинформационные системы. Геоинформатика в литературоведении. Использование открытых электронных картографических инструментов

4 часа аудиторной работы.

Foote, Lynch, 2015; Журкин, Шайтура, 2009; Moretti, 2005;

Тема 7. Цифровые научные (критические) издания художественной литературы.

Классическое наследие в цифровую эпоху. Новые возможности для отображения традиционного критического аппарата. Типология существующих критических (научных) цифровых изданий.

Международный стандарт текстологической и метатекстовой разметки TEI (Text Encoding Initiative). Кодирование информации о тексте при помощи TEI. Цифровые издания на базе TEI. Инструменты для публикации TEI в цифровой среде.

TEI P5 Guidelines; Robinson, 2013; Schreibman, 2013

12 часов аудиторной работы.

Тема 8. Культуромика (количественные методы исследования культуры). «Большие данные» в гуманитарных науках. Визуализация данных

«Большие данные» в гуманитарных науках; выявление трендов; анализ данных Википедии и Twitter; количественные исследования кино, театра, креативных индустрий. Количественные исследования медиа. Глобальные и локальные тренды. Работа с массивами неструктурированных текстов из блогов и социальных сетей. Семантический веб и связанные открытые данные (linked open data). Общедоступные базы данных (DBPedia, FreeBase, FOAF) и работа с ними; онтологии (SOWA, Dublin Core) и их применение.

20 часов аудиторной работы.

Michel *et al*, 2011; Lieberman *et al*, 2007; Hoover *et al*, 2014; Schreibman *et al*, 2004;

Образовательные технологии

На лекциях используются презентации, аудиозаписи и фрагменты видео. В ходе практических занятий применяются современные средства компьютерного анализа данных (соответствующие библиотеки языков программирования R, Python, пакет Mallet) и их визуализации (Gephi).

Оценочные средства для текущего контроля и аттестации студента

Итоговый экзамен по дисциплине проводится в форме защиты проекта (proof of concept) цифрового гуманитарного исследования или ресурса. Соответственно, от студентов ожидается подготовка такого проекта в качестве текущей работы по курсу; в процессе защиты студент должен показать глубокое знакомство с теми из основных современных методов гуманитарных исследований, которые применяются в её/его проекте.

Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на семинарских занятиях: активность студентов в дискуссиях, работу с рекомендуемой литературой, доклады на выбранные темы. Оценки за работу на семинарских и практических занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за работу на семинарских и практических занятиях определяется перед промежуточным или итоговым контролем - $O_{аудиторная}$.

Освоение части тем предполагает выполнение домашних заданий. Накопленная оценка по 10-ти балльной шкале за домашние задания определяется перед промежуточным или итоговым контролем - $O_{домашняя}$.

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где $O_{экзамен}$ – оценка за работу непосредственно на экзамене:

$$O_{итоговый} = 0{,}4 \cdot O_{экзамен} + 0{,}3 \cdot O_{аудиторная} + 0{,}3 \cdot O_{домашняя}$$

Способ округления накопленной оценки итогового контроля в форме экзамена: в пользу студента.

Учебно-методическое и информационное обеспечение дисциплины

1.1 Базовый учебник

Schreibman S., Siemens R., Unsworth, J. (eds.) (2004) Companion to Digital Humanities. Blackwell Companions to Literature and Culture. Oxford: Blackwell.

1.2 Основная литература

Hoover, D. Culpeper, J., O'Halloran, K. (2014) Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. Routledge Advances in Corpus Linguistics.
Moretti, F. (2013) Distant Reading. London: Verso.
Моретти Ф. Дальнее чтение. М., Издательство Института Гайдара, 2016.

1.3 Дополнительная литература

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
Blei D. (2012) Introduction to Probabilistic Topic Models // Communications of the ACM. — С. 77–84.
Burrows, J. (2002) 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17:267-87
Eder, M. Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: "Digital Humanities 2013: Conference Abstracts". University of Nebraska-Lincoln, Lincoln, NE, pp. 487-89.
Eder M. (2011) Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. Studies in Polish Linguistics, 6:99–114.
Foote K., Lynch M., (2015) Geographic Information Systems as an Integrating Technology: Context, Concepts, and Definitions". The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder.
Hoover, D. (2007) The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory. Digital Humanities Quarterly 1.2.
Jockers M. (2013) Macroanalysis: Digital Methods and Literary History. Champaign, IL: University of Illinois Press.
Juola, P. (2006). Authorship Attribution. Foundations and Trends in Information Retrieval
Juola, P., Baayen, H. (2005) A Controlled-corpus Experiment in Authorship Identification by Cross-Entropy, Literary and Linguistic Computing 20 (Suppl 1), pp. 59-67

Kestemont, M. (2014) Function words in authorship attribution. from black magic to theory? In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pages 59–66, Gothenburg, Sweden.
Lieberman E., Michel, J.B., Jackson J., Tang T., Nowak M. (2007) Quantifying the Evolutionary Dynamics of Language. Nature. p. 449.
Liu, A. (2011) The State of the Digital Humanities: A Report and a Critique. Arts and Humanities in Higher Education 2.1-2. pp. 8-41.
Mendenhall T. (1887) The Characteristic Curves Of Composition. Science : Vol. 9, Issue 214S, pp. 237-246
Michel, J.B., (2011). Quantitative analysis of culture using millions of digitized books . Science, 331(6014): pp. 176–82.
Moretti F. (2005) Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.
Moretti F. (2011) Network Theory, Plot Analysis. New Left Review 68, March-April
Mosteller F., Wallace D., (1963) Inference in an Authorship Problem. Journal of the American Statistical Association, Volume 58, Issue 302, pp. 275 - 309.
Newman M. (2010) Networks: An Introduction. Oxford: Oxford University Press.
Orekhov B. (2015) Bashkir poetic corpus as a DH-resource / NRU HSE. Series WP BRP "Linguistics".
Robinson P. (2013) Towards a Theory of Digital Editions. Variants: The Journal of the European Society for Textual Schola. Vol. 10, p. 105
Schreibman S. (2013) Digital Scholarly Editing. Literary Studies in the Digital Age: An Evolving Anthology// https://dlsanthology.mla.hcommons.org/digital-scholarly-editing/
Schreibman S., Siemens R. (eds.) (2008) A Companion to Digital Literary Studies. Oxford: Blackwell.
Воронцов К.В. Вероятностное тематическое моделирование // http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf, 2013.
Журкин И. Г., Шайтура С. В. Геоинформационные системы. — Москва: Кудиц-пресс, 2009.
Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. М.: Языки славянской культуры, 2016.
Мухин М. Ю. Лексическая статистика и идиостиль автора: корпусное идеографическое исследование : на материале произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова: диссертация на соискание ученой степени доктора филологических наук. [Место защиты: ГОУВПО "Уральский государственный университет"]. - Екатеринбург, 2011. - 383 с.
Ярхо Б. И., Методология точного литературоведения: Избранные труды по теории литературы, Ред. М. В. Акимова, И. А. Пильщиков и М. И. Шапир, Под общей редакцией М. И. Шапира, Москва: Языки славянских культур, 2006, xxxii, 927 с.