Программа учебной дисциплины «Цифровые методы для гуманитариев»

Утверждена
Академическим советом ОП[1]
Протокол № от _____

Разработчик	Куприянов Алексей Валерьевич, доцент департамента социологии; Маслинский Кирилл Александрович, доцент департамента социологии
Число кредитов	3
Контактная работа (час.)	42
Самостоятельная работа (час.)	72
Курс, Образовательная программа	Образовательная программа «Филология»
Формат изучения дисциплины	С использованием онлайн-курса

1. **Цель, результаты освоения дисциплины и пререквизиты**

Цель дисциплины «Цифровые методы для гуманитариев» — познакомить студентов-филологов с базовыми понятиями и методами анализа и визуализации данных, а также дать базовые навыки работы с программным инструментарием, необходимым для анализа и визуализации данных на примере статистического пакета R. Содержание курса охватывает: основы статистики, визуализацию данных, основы количественного анализа текстов и основы программирования на R. Освоение курса поможет студентам освоить основные категории и инструменты, необходимые для выполнения количественного анализа данных в гуманитарных науках, а также заложит основу для дальнейшего обучения современным методам анализа и визуализации данных.

Формат изучение дисциплины предполагает самостоятельное освоение студентами части учебного материала в формате онлайн-курса (blended-learning).

Изучение дисциплины ««Ключевые тексты русской литературы»» обеспечивает формирование у выпускников бакалавриата по направлению подготовки 45.04.01 «Филология» следующих компетенций:

- УК-2 — Способен выявлять научную сущность проблем в профессиональной области.

[1] Для *ПУД из общеуниверситетского пула – Руководитель Департамента.*

- ПК6 — Способен проводить научные исследования в конкретной области филологического знания с формулировкой аргументированных умозаключений и выводов.
- ПК9 — Способен участвовать в научных дискуссиях, выступать с сообщениями и докладами, представлять материалы собственных исследований в устной и письменной форме, в том числе с использованием компьютерных технологий.
- ПК17 — Способен участвовать в последовательной реализации индивидуального или коллективного проекта.

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах**[1]	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
	лк см onl/cp		
Тема 1. Введение. Обзор инструментов	1 4 12	Студент имеет представление о наиболее общих программных инструментах, используемых в цифровой гуманиоре, знает основы языка регулярных выражений, умеет использовать их в программных скриптах и владеет основами пользования imagemagick и OCR-Tesseract для подготовки текстов к анализу.	Практическая работа на семинарах, самостоятельная работа над итоговым домашним заданием.
Тема 2. Данные	2 2	Студент имеет представления о концепции tidy data и case-variable structure, представлении данных в форматах текста с разделителями, XML и JSON.	Практическая работа на семинарах по смежным темам

[1] Не заполняется для ПУД, которые не вошли в УП ОП и не запланированы в расписании учебных занятий

Тема 3. Визуализация паттернов и формальные методы анализа	5 6 16	Студент знает основы прикладной статистики, умеет выполнять стандартные задания по визуализации данных и формальному анализу данных (получение дескриптивных статистик, расчет корреляций, линейной регрессии, сравнение двух и более выборок, анализ таблиц сопряженности в среде статистического программирования и анализа данных R)	Практическая работа на семинарах, экзамен.
Тема 4. Элементы креативной инфографики	2 2 6	Студент имеет представление о принципах креативной инфографики и умеет реализовывать свои идеи в области креативной инфографики средствами R.	Самостоятельная работа на семинаре.
Тема 5. Лексическая статистика	2 2 9	Студент знает принцип распределения языковых единиц в текстах и понимает его следствия для теоретических и прикладных вопросов количественного анализа текста	Практическая работа на семинарах, экзамен
Тема 6. Классификация текстов	2 4 9	Студент имеет представление о принципе решения задач классификации в машинном обучении, умеет применять наивный байесовский классификатор	Практическая работа на семинарах, самостоятельная работа над домашним заданием
Тема 7. Дистрибутивная семантика	4 2 9	Студент знаком с содержанием дистрибутивной гипотезы и современными данными по этой проблеме, имеет представление о сфере применения	Практическая работа на семинарах, экзамен

		дистрибутивных методов в задачах количественного анализа текста	
Тема 8. Тематическое моделирование	2 2 9	Студент имеет представление о логике работы и сфере применения методов тематического моделирования	Практическая работа на семинаре, самостоятельная работа над домашним заданием
Часов по видам учебных занятий:	20		
	22		
	72		
Итого часов:	114		

Формы учебных занятий:
лк – лекции в аудитории;
см - семинары/ практические занятия/ лабораторные работы в аудитории;
onl – лекции или иные виды работы студента с помощью онлайн-курса;
ср – самостоятельная работа студента.

Содержание разделов дисциплины

Тема 1. Введение. Обзор инструментов. (1.1) Вступление: о важности визуализации паттернов. Квартет Энскомба. Задачи анализа данных: описание разнообразия и поиск взаимосвязей. (1.2) Обзор основных инструментов, изучаемых в рамках курса и их место в задачах анализа данных. Редакторы кода. Язык регулярных выражений. Среда статистического программирования и анализа данных R. Imagemagick. OCR-Tesseract. QGIS.

Тема 2. Данные. Данные и метаданные, концепция tidy data. Case-variable структра, агрегированные и дезагрегированные данные. Классификация переменных и шкал. Специфика цифрового представления данных. Кодировки текстовых файлов и обработка концов строк. Delimited text, XML, JSON.

Тема 3. Визуализация паттернов и формальные методы анализа. (3.1) Отображение разнообразия: гистограммы и столбчатые диаграммы. Графические образы моделей с двумя переменными: диаграмма рассеяния, диаграммы разброса (множественный

boxplot), диаграмма рассеяния с добавленным шумом, структурированные столбчатые диаграммы. (3.2) Меры центральности и разброса, их особенности. Асимметрия и эксцесс. Основные представления о нормальном распределении. (3.3) Выборочный метод. Точечные и интервальные оценки параметров. Статистическая гипотеза и ее тестирование, p-value. (3.4) Связь двух количественных переменных. Корреляция. Основные представления о линейной регрессии. Сравнение двух и более групп между собой. Параметрические и непараметрические методы. Общее понятие об обобщенной линейной модели. Анализ таблиц сопряженности.

Тема 4. Элементы креативной инфографики. Использование инфографики в просопографических проектах, проектах по Distant reading, карты и социальные сети.

Тема 5. Лексическая статистика. Частотное распределение лексики в языке. Закон Ципфа. Доля hapax legomena. Скорость роста словаря. Меры лексического разнообразия и их применимость. Распределение лексики в текстах коллекции. Взвешенная частотность. TF-IDF. Прочие меры лексической дисперсии. Коллокации. Формальные определения и лингвистический смысл коллокаций. Меры ассоциации. Коэффициент взаимной информации (MI). Извлечение ключевых слов. Метод контрастного корпуса. Отношение правдоподобия. Диахронический анализ лексической частотности.

Тема 6. Классификация текстов. Задача классификаци в машинном обучении. Векторное представление текста для задач информационного поиска. Открытые и закрытые классы слов. Стоп-слова. Динамические списки стоп слов. Порог отсечения по частотности и DF. Классификация текстов. Теорема байеса. Популярные алгоритмы классификации: наивный байесовский метод, метод опорных векторов, деревья принятия решений.

Тема 7. Дистрибутивная семантика. Дистрибутивная семантика. Совместная встречаемость и семантическая близость. Пространственное моделирование семантических отношений (word space). Методы снижения размерности векторных пространств. Латентный семантический анализ. Векторные представления дистрибуции слова в пространствах низкой размерности (word embeddings).

Тема 8. Тематическое моделирование. Операционализация понятия «тема» как вероятностного распределения лексики. Латентное размещение Дирихле (LDA). Процедура тематического моделирования. Препроцессинг. Сегментация текстов.

Сэмплирование Гиббса. Интерпретация тем. Оценка качества модели. Использование результатов тематического моделирования в задаче классификации текстов. Оценка качества классификации (продолжение). Таблица сопряженности. Точность, полнота, F-мера. Матрица неточностей. Каппа-статистика.

3. Оценивание

3.1. Формула результирующей оценки

$$О_{рез} = 0,3(О_{дз1} + О_{дз2}) +0,4О_{экз}$$

$О_{рез}$ – результирующая оценка;
$О_{дз1}$ — домашнее задание (темы 1—4);
$О_{дз2}$ — домашнее задание (темы 5—8);
$О_{экз}$ — итоговый экзамен.

3.2. Критерии оценивания

Домашнее задание представляет собой письменный отчет по результату выполнения двух лабораторных работ, предлагаемых в курсе, либо письменный отчет по результату проведения индивидуального проекта по анализу текстовой коллекции. Обе формы контроля являются эквивалентными. Каждая лабораторная работа оценивается индивидуально, результирующая оценка за домашнее задание — среднее арифметическое оценок за отдельные лабораторные работы. За каждую не сданную работу из результирующей оценки вычитается 1 балл. За индивидуальный проект выставляется одна итоговая оценка.

Экзаменационная оценка выставляется за развернутый письменный ответ по нескольким вопросам из тем, рассмотренных в ходе курса. В ответе необходимо продемонстрировать понимание смысла темы, должны быть использованы актуальные и адекватные научные источники. Плагиат в ответах недопустим.

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

В случае, если студент не имеет возможности сдать письменную работу или экзамен в срок по уважительной причине, по согласованию с преподавателем назначается дополнительный день сдачи экзамена или продлевается срок сдачи письменной работы.

4. Примеры оценочных средств

Примеры самостоятельных домашних заданий:

1. Выберите 20 биографий из словаря русских писателей, постройте парсер, помогающий извлечь максимум биографической информации в машиночитаемом виде. Результаты работы сохраните в виде текста с разделителями. Сдайте исходный текст, код парсера и данные в виде архива.
2. Выберите для распознавания фрагмент книги в русской дореформенной орфографии приблизительным объемом 20 тыс. символов (приблизительно 10 страниц по 30 строк по 66 символов в строке или эквивалентный, учитывая особенности верстки). Произведите тренировку OCR-Tesseract на его основе. Сдайте все необходимые для тренировки файлы и итоговую модель в виде архива.

Примеры практических заданий для выполнения на семинаре:

1. Выполните преобразование изображения при помощи программ jhead и imagemagick, подготовив фотографию страницы книги к распознаванию, произведите распознавание при помощи программы tesseract.
2. Постройте стандартную визуализацию для одной переменной или для отображения взаимосвяязи двух переменных, учитывая характер переменных и шкал.
3. Используя цикл, постройте серию однотипных визуализаций для подмножеств тренировочного набора данных.
4. Исходя из характера двух переменных подберите адекватный метод анализа их возможной взаимосвязи.
5. Сравните два подмножества тренировочного набора данных по одной из количественных переменных. Подберите адекватный метод сравнения, учитывая характер распределения значений переменной и предложите интерпретацию результата.
6. Проанализируйте таблицу сопряженности между двумя качественными переменными. При необходимости произведите перегруппировку данных. Предложите интерпретацию результата.
7. Средствами R постройте карту, отражающую количественные характеристики картируемых объектов. То же для качественных характеристик.

Примерное задание к экзаменационной контрольной работе:

1. На основе тренировочного набора данных постройте необходимые визуализации. Протестируйте гипотезы о наличии связей между переменными (не менее трех видов анализа, сообразно характеру переменных), дайте интерпретацию результатов.

5. Ресурсы

5.1 Основная литература

1. *Кабаков Р. И.* R в действии: анализ и визуализация данных в программе R / Пер. с англ. П. А. Волковой. М.: ДМК Пресс, 2014. 587 с.

5.2 Рекомендуемая дополнительная литература

1. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. Т. 23.

2. Bamman D., Eisenstein J., Schnoebelen T. Gender identity and lexical variation in social media // Journal of Sociolinguistics. 2014. T. 18, No 2. C. 135—160.
3. Jockers M. L., Mimno D. Significant themes in 19th-century literature // Poetics. 2013. T. 41, No 6. C. 750—769.

5.3 Программное обеспечение

Для успешного освоения дисциплины, студент использует следующие программные продукты:

- При работе с операционными системами MS Windows:
  - MS Office (Word, Excel) (из внутренней сети университета (договор)) или Open/Libre Office (Writer, Calc) (свободно распространяемый продукт)
  - Редактор кода для MS Windows: Notepad++ https://notepad-plus-plus.org/ (свободно распространяемый продукт)
  - Интерпретатор Strawberry Perl http://strawberryperl.com/ (свободно распространяемый продукт)
- При работе с MacOS X:
  - Open/Libre Office (Writer, Calc) (свободно распространяемый продукт)
  - Редактор кода Atom https://flight-manual.atom.io/getting-started/sections/installing-atom/#platform-mac (свободно распространяемый продукт)
  - Интерпретатор Perl (проверить наличие предустановленного дистрибутива)
- Кросс-платформенные свободно распространяемые продукты:
  - Редактор изображений imagemagick http://www.imagemagick.org/script/download.php
  - Редактор заголовков JPEG файлов jhead https://www.techworld.com/download/audio-video-photo/jhead-30-3330267/
  - Система распознавания текста ocr-tesseract https://github.com/tesseract-ocr/tesseract
  - Утилита тренировки ocr-tesseract jTessBoxEditor http://vietocr.sourceforge.net/training.html
  - Среда статистического программирования и анализа данных R https://www.r-project.org/

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

1. Тренировочные наборы данных и код https://github.com/alexei-kouprianov/Breaking-the-ice-with-R
2. Тренировочный набор данных: Контурные карты Российской Империи и сопредельных стран на 1905 год https://github.com/alexei-kouprianov/GIS.projects/tree/master/Marx.1905
3. Каталог лингвистических ресурсов для обработки русского языка — http://nlpub.ru.

5.5 Материально-техническое обеспечение дисциплины

На лекционных занятиях необходим компьютер, проектор для презентаций, аудиоколонки. На семинарских занятиях необходимы компьютеры с доступом в Интернет с установленными программами.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

6.1.2. *для лиц с нарушениями слуха*: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

6.1.3. *для лиц с нарушениями опорно-двигательного аппарата*: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.

7. Дополнительные сведения

По желанию разработчика в ПУД могут быть включены другие содержательные элементы, например, методические рекомендации для студента и преподавателя, описание применяемых образовательных технологий