Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Институт общественных наук
(наименование института)
Кафедра истории экономики
(наименование кафедры)

УТВЕРЖДЕНА
кафедрой истории экономики
Протокол от «29» августа 2016 г.
№ 3

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.В.ДВ.17.2 Компьютерные методы анализа текста
(индекс и наименование дисциплины)

46.03.01 История
(код и наименование направления подготовки)

История государства и власти (Liberal Arts)
направленность (профиль)

бакалавр
(квалификация)

очная
(форма обучения)

Год набора - 2017

Москва, 2016 г.

Автор–составитель:

к.и.н., доцент	истории экономики	Кончаков Р.Б.
(ученое звание, ученая степень, *должность)*	*(наименование кафедры)*	*(Ф.И.О.)*

Заведующий кафедрой
истории экономики, к.и.н., доцент__ Кончаков Р.Б.
*(наименование кафедры) (*ученое звание, ученая степень,*)* *(Ф.И.О.)*

СОДЕРЖАНИЕ

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы……
2. Объем и место дисциплины в структуре образовательной программы……...
3. Содержание и структура дисциплины……....
4. Материалы текущего контроля успеваемости обучающихся и фонд оценочных средств промежуточной аттестации по дисциплине ……...
5. Методические указания для обучающихся по освоению дисциплины ……...
6. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине …….
   6.1. Основная литература……..……...……
   6.2. Дополнительная литература……...……..……
   6.3. Учебно-методическое обеспечение самостоятельной работы…….
   6.4. Нормативные правовые документы…….……..……..
   6.5. Интернет-ресурсы…….……..……...
   6.6. Иные рекомендуемые источники…….……..
7. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы ……...…..……..

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения программы

1.1. Дисциплина Б1.В.ДВ.17.2 «Компьютерные методы анализа текста» обеспечивает овладение следующими компетенциями с учетом этапа:

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенций
СК ОС LA- 12	Способность использовать современное программное обеспечение и электронные ресурсы в планировании и реализации гуманитарных цифровых проектов	СК ОС LA– 12.1	Владение основными методическими принципами использования программного обеспечения и информационными ресурсами в гуманитарном исследовании

1.2. В результате освоения дисциплины у студентов должны быть сформированы:

Код этапа освоения компетенции	Результаты обучения
СК ОС LA - 12.1	на уровне знаний: сформированы знания принципов проведения гуманитарного исследования с привлечением программно-технических средств; в том числе знание терминологии, концептуальных подходов, теорий и тенденций в изучении методов анализа текстов документов; методов анализа текстов документов; алгоритма применения методов анализа текстов документов.
	на уровне умений: сформированы умения использовать специализированные программные продукты для анализа текстов, в том числе умение характеризовать и интерпретировать приемы, правила и процедуры различных видов анализа текстов документов оценить корректность применения того или иного метода для анализа данного вида документов оценивать различные методы анализа документов и выбирать оптимальные методы критически воспринимать и анализировать содержание документа.
	на уровне навыков: сформированы навыки планирования и реализации цифровых гуманитарных проектов в части интерпретации теоретических основ различных методик анализа документов; корректного применения инструментария специальных методов анализа; выработки рекомендаций по применению методов анализа текстов документов в документоведении.

2. Объем и место дисциплины в структуре ОП ВО

Объем дисциплины

В соответствии с учебным планом дисциплина Б1.В.ДВ.17.2 «Компьютерные методы анализа текста» входит в состав дисциплин по выбору вариативной части блока Б1 «Дисциплины» и изучается в 5 семестре. Общая трудоемкость дисциплины составляет 72 часа (2 з.е.)

Количество академических/астрономических часов, выделенных на контактную работу с преподавателем – 28/21 часов, на самостоятельную работу обучающихся – 44/33 часа.

Место дисциплины в структуре ОП ВО

Содержание данной дисциплины опирается на ранее изученную дисциплину Б1.Б.12 «Информатика», которая относится к дисциплинам базовой части блока Б1 и изучается в 1 семестре.

Дисциплина реализуется после изучения базовой части программы.

3. Содержание и структура дисциплины

Таблица 1.

№ п/п	Наименование тем (разделов)	Объем дисциплины, час.						Форма текущего контроля успеваемости[4], промежуточной аттестации
		Всего	Контактная работа обучающихся с преподавателем по видам учебных занятий				СР	
			Л	ЛР	ПЗ	КСР		
Очная форма обучения								
Тема 1	Методика анализа документов: основные понятия	8/6			4/3		4/3	Домашнее задание
Тема 2	Эволюция методологических основ методики анализа текстов документов	12/9			4/3		8/6	Домашнее задание
Тема 3	Семиотика	12/9			4/3		8/6	Домашнее задание
Тема 4	Контент-анализ документов	12/9			4/3		8/6	Домашнее задание
Тема 5	Классификация документов	14/10,5			6/4,5		8/6	Домашнее задание
Тема 6	Тематическое моделирование	14/10,5			6/4,5		8/6	Домашнее задание
Промежуточная аттестация								**Зачет с оценкой**
Всего:		**72/54**			**28/21**		**44/33**	

Содержание дисциплины

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)
Тема 1	Методика анализа документов: основные понятия	Понятие научного метода и научной методики. Связь метода с теоретическими и мировоззренческими аспектами познания. Структура метода. Теория и методика источниковедения как методологическая основа разработки методов анализа документов. Исторический источник и документ: соотношение понятий. Исторический источник/документ как явление культуры.

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)
Тема 2	Эволюция методологических основ методики анализа текстов документов	Значение рационалистической научной парадигмы в становлении критического метода анализа документов. Вклад позитивизма в развитие методов анализа. Зарождение междисциплинарной кооперации социогуманитарных дисциплин и ее влияние на развитие методики анализа. Роль герменевтики, структурной лингвистики, семиотики в понимании природы текста. Значение «школы Анналов» в разработке проблемы методов изучения документов. Соотношение традиционных, междисциплинарных и специальных методов анализа документов.
Тема 3	Семиотика	Понятие семиотики. Семиотика как метод анализа культурного контекста человеческого существования. Культурно-семиотический подход к исследованию социальной реальности. Культура как текст. Знаки, символы, культурные коды в системе культуры. Вклад отечественных и зарубежных исследователей в развитие семиотики. Р. Барт. Ю.М. Лотман и московско-тартуская семиотическая школа. Историческая семиотика. Текст исторического источника как зашифрованное описание. Семиотические методы анализа текстов исторических источников и документов: возможности и ограничения.
Тема 4	Контент-анализ документов	Понятие контент-анализа. Технология проведения контент-анализа массовых и нарративных документов. Качественный и количественный анализ текста документа при проведении контент-анализа. Три стадии контент-анализа. Использование возможностей контент-анализа в документационном обеспечении управления, интерпретации содержания документов путем выявления и анализа скрытых связей между их смысловыми концептами

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)
Тема 5	Классификация документов	Векторная модель документа. Матрица терминов – документов. Взвешивание терминов: нормализация по длине документа, TF-IDF. Проблема разреженных данных. Методы снижения размерности. Стоп-слова. Отбор значимых свойств (feature selection). Задача машинного обучения. Машинное обучение с учителем. Обучающая и тестовая выборки. Алгоритм обучения. Задача классификации текстов. Области применения классификации в обработке естественного языка. Оценка качества классификации. Точность. Кросс-валидация. Понятие корпус. Лингвистическая аннотация. Иерархия языковых уровней. Лексика. Частотный анализ текстов. Закон Ципфа. Открытые и закрытые классы слов. Морфологический анализ. Части речи. Стемминг и лемматизация. Полный и частичный синтаксический анализ. N-граммы.
Тема 6	Тематическое моделирование	Дистрибутивная гипотеза в семантике. Латентный семантический анализ. Вероятностный латентный семантический анализ (pLSA). 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.

4. Материалы текущего контроля и фонд оценочных средств промежуточной аттестации по дисциплине

4.1. Формы и методы текущего контроля успеваемости и промежуточной аттестации.

4.1.1. В ходе реализации дисциплины «Компьютерные методы анализа текста» используются следующие методы текущего контроля и успеваемости обучающихся:
– при проведении практических занятий:
опрос,
обсуждение домашних заданий.

4.1.2. Зачет проводится с применением следующих форм (средств):
Промежуточная аттестация проводится в форме устного зачета, предполагающего ответы на поставленные вопросы.

4.2. Материалы текущего контроля успеваемости.
В процессе преподавания данной дисциплины используются как классические методы

обучения (семинары), так и различные виды самостоятельной работы студентов по заданию преподавателя, которые направлены на развитие творческих качеств студентов и на поощрение их интеллектуальных инициатив.
В рамках данного курса используются такие активные формы обучения, как:
• выполнение промежуточных тестов по итогам семинарских занятий.

Вопросы для самостоятельной подготовки по темам дисциплины:
1. Стилометрия. История дисциплины и классические результаты.
2. Алгоритмы классификации. Наивный Байес.
3. Алгоритмы классификации. Деревья принятия решений.
4. Алгоритмы классификации. Support vector machine (SVM).
5. Проблема переобучения (overfitting) и методы ее решения.
6. Обзор разновидностей тематических моделей. Twitter-LDA. Author-LDA. Диахронические модели.
7. Методы оценки качества тематических моделей. Perplexity. PMI.
8. Метрики качества отдельных тем.
9. Иерархические тематические модели. Pachinko allocation.
10. Тематическая кластеризация текстов.
11. Обзор работ по анализу тональности текстов на русском языке.
12. Словарь оценочной лексики для области товаров Четверкина. Методология составления.
13. Коллокации. Методы обнаружения коллокаций.
14. Сравнение методов выделения характерной лексики.
15. Обзор работ по извлечению именованных сущностей из текстов на русском языке.

4.3. Оценочные средства для промежуточной аттестации.

4.3.1. Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Показатели и критерии оценивания компетенций с учетом этапа их формирования

Код компетенции	Наименование компетенции	Код этапа освоения компетенции	Наименование этапа освоения компетенций
СК ОС LA- 12	Использовать современное программное обеспечение и электронных ресурсов в планировании и реализации гуманитарных цифровых проектов.	СК ОС LA– 12.1	Владение основными методическими принципами использования программного обеспечения и информационными ресурсами в гуманитарном исследовании

Этап освоения компетенции	**Показатели оценивания**	**Критерии оценивания**
СК ОС LA -12.1. Владение основными методическими	Понимание возможностей и пределов использования программного обеспечения	Студент понимает аналитические возможности различных программных продуктов, умеет

принципами использования программного обеспечения и информационными ресурсами в гуманитарном исследовании	и информационных систем в гуманитарном исследовании; Понимание методологических принципов выбора инструментов и методов для гуманитарного исследования и представления результатов.	выбирать наиболее эффективные программные средства и информационные системы для реализации гуманитарного исследования, владеет широким инструментарием для проведения исследования, может комбинировать различные программные продукты для достижения наилучшего результата.

4.3.2 Типовые оценочные средства

Задание к зачету предполагает устные ответы на поставленные вопросы.

Примерные вопросы к устному зачету:

1. Понятие метода исследования. Теоретический и практический аспекты метода.
2. Понятия «документ», «источник», их соотношение.
3. Методы исследования текстов документов, их классификация.
4. Вклад позитивизма в разработку методики анализа текстов документов.
5. Междисциплинарные методы: понятие, классификация. Место междисциплинарных методов в системе методов анализа документов.
6. Феномен междисциплинарности, его становление, влияние на современную методику анализа документов.
7. Герменевтический и феноменологический подходы к интерпретации текстов, их значение для методики анализа документов.
8. Методы анализа текстов культуры в современной антропологии, возможности их применения.
9. Методика анализа документов в ситуации постструктуралистского/постмодернистского «вызова»: возможности и ограничения.
10. Семиотика. Метод семиотического анализа текстов и его применение.
11. Дискурсивный анализ текстов документов.
12. Метод контент-анализа и его применение в исследовании документов.
13. Филологические методы исследования текстов: концептуальный и денотативный анализ.
14. Информативно-целевой анализ документов.
15. Методы психоанализа в исследовании текстов документов.
16. Формализованные методы анализа документов.

Шкала оценивания.

Форма промежуточной аттестации	Критерии оценивания	Оценка

Зачет с оценкой	- Студент понимает аналитические возможности различных программных продуктов; - понимает принципы выбора наиболее эффективных программных средств и информационных систем для реализации гуманитарного исследования; - владеет широким инструментарием для проведения исследования, может комбинировать различные программные продукты для достижения наилучшего результата.	81–100 баллов Отлично
	- Студент понимает аналитические возможности различных программных продуктов; - умеет отбирать программные средства и информационные системы для реализации гуманитарного исследования, но не всегда действует наиболее эффективным образом; - владеет определённым инструментарием для проведения исследования, комбинирует некоторые программные продукты для достижения результата.	61–80 баллов Хорошо
	- Студент частично понимает аналитические возможности различных программных продуктов; - выбирает наиболее очевидные программные средства и информационные системы для реализации гуманитарного исследования, при этом не учитываются особенности решений, не достигается эффективной реализации проекта; - владеет некоторым инструментарием для проведения исследования, не может комбинировать программные продукты для достижения результата.	41–60 баллов Удовлетворительно
	- Студент не понимает аналитические возможности различных программных продуктов; - не понимает принципов отбора инструментов для реализации гуманитарных проектов; - не владеет инструментарием для проведения исследования.	40 и менее неудовлетворительно

4.4. Методические материалы

В процессе преподавания данной дисциплины используются как классические методы обучения (практические занятия), так и различные виды самостоятельной работы студентов по заданию преподавателя, которые направлены на развитие творческих качеств студентов и на поощрение их интеллектуальных инициатив.

5. Методические указания для обучающихся по освоению дисциплины

Главные задачи курса с одной стороны, познакомить студентов с результатами, достигнутыми в области обработки естественного языка, а с другой, — стимулировать и подготовить их к аналитической работе с массивами текстовых данных в теоретических и прикладных социологических исследованиях.

Объем курса и его место в образовательной программе, в которой отсутствуют базовые лингвистические курсы, а курсы по программированию в лучшем случае являются факультативными, не позволяют дать систематическое изложение всех разделов и методов автоматической обработки языка и компьютерной лингвистики.

В то же время, задачи курса предполагают возможность для слушателей пройти путь от теоретического обсуждения методов работы с текстом к их практическому применению. Поэтому в качестве основы для построения курса выбран принцип разбора кейсов — нескольких современных исследований, в которых проводился анализ большого объема текстовых данных. В рамках курса подробно обсуждаются теоретические основания, методология и программный инструментарий, необходимые для проведения аналогичных исследований.

На практических занятиях и в ходе самостоятельной работы по курсу слушатели получают возможность применить изученные методы к предложенным в рамках курса или к их собственным текстовым коллекциям.

Для оценивания уровня знаний, умений и навыков студентов используется комплекс контрольных мероприятий текущих и итоговых.

Текущие мероприятия включают:

1. Семиотический анализ текста: теоретический и практический аспект (на примере анализа документа)
2. Дискурсивный анализ в исторических исследованиях: интерпретация опыта применения.
3. Контент-анализ документов: алгоритм и практика применения (на примере исследования массовых документов)

Технология организации самостоятельной работы обучающихся включает использование информационных и материально-технических ресурсов образовательного учреждения.

Для подготовки к практическому занятию необходимо:

- внимательно прочесть материал относящихся к теме семинарского занятия;
- ознакомится с дополнительными источниками по тематике семинарского занятия;
- ответить на контрольные вопросы к семинарским занятиям (необходимо подготовить развернутый ответ на каждый из поставленных вопросов);
- зафиксировать для себя, в какой части материал семинарского занятия остался для вас непонятен;
- подготовить вопросы для преподавателя по непонятой части тематики семинарского занятия;
- постараться получить ответы от преподавателя заранее (на консультации вперед занятием).

Готовиться можно индивидуально, парами или в составе малой группы, последние являются эффективными формами работы.

В разделе 6 (п. 6.1., п. 6.2.) указан перечень основной и дополнительной литературы, который рекомендуется обучающимся при подготовке к семинарским занятиям и выполнении самостоятельной работы.

6. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", включая перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Основная литература.

1. Методы анализа предметных областей .Кемерово: Кемеровский государственный институт культуры.2011.http://www.iprbookshop.ru/29684.

6.2. Дополнительная литература.

1. Изюмов А.А., Коцубинский В.П. Компьютерные технологии в науке и образовании : Учебное пособие .Томск: Эль Контент,2012.http://www.iprbookshop.ru/13885.html

6.3. Учебно-методическое обеспечение самостоятельной работы.

Положение об организации самостоятельной работы студентов федерального государственного бюджетного образовательного учреждения высшего образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации» (в ред. приказа РАНХиГС от 11.05.2016 г. № 01-2211). http://www.ranepa.ru/images/docs/prikazy-ranhigs/Pologenie_o_samostoyatelnoi_rabote.pdf

6.4. Нормативные правовые документы.

1. Федеральный закон от 27 июля 2006 года № 149-ФЗ «Об информации, информационных технологиях и защите информации» (в ред. ФЗ от 07.06.2017 N 109-ФЗ) // Справочно-правовая система Консультант+ (дата обращения: 15.06.2017).

6.5. Интернет-ресурсы, справочные системы.

1. http://www.dialog-21.ru/ — Диалог.Международная конференция по компьютерной лингвистике.
2. http://nlpub.ru — Каталог лингвистических ресурсов для обработки русского языка. http://www.regular-expressions.info — The Premier website about Regular Expressions.
3. http://sentiment.christopherpotts.net/ — Sentiment symposium tutorial.
4. ttp://www.aclweb.org/anthology/ — ACL Anthology A Digital Archive of Research Papers in Computational Linguistics.

6.6. Иные рекомендуемые источники.

1. Мазур Л.Н. Методы исторического исследования Екатеринбург, 2010
2. Поршнева О.С. Междисциплинарные методы в историко-антропологических исследованиях. Екатеринбург: УрГУ, 2005.
3. Поршнева О.С. Междисциплинарные методы в историко-антропологических исследованиях. Изд. 2-е, доп.: Учебное пособие для вузов (гриф УМО). Екатеринбург: УГТУ-УПИ, 2009.
4. Барт Р. Избранные работы: Семиотика. Поэтика: Пер. с фр. М., 1994.
5. Бородкин Л.И. Многомерный статистический анализ в исторических исследованиях. М., 1986.
6. Герменевтика в России. Сб. ст. Воронеж, 2002. Вып. 1.
7. Источниковедение: Теория. История. Метод. Источники российской истории: Учеб.

пособие / И. Н. Данилевский, В. В. Кабанов, О. М. Медушевская и др. М., 2000.
8. Источниковедение новейшей истории России: теория, методология, практика: Учебник / А.К. Соколов, Ю.П. Бокарев, Л.В. Борисова и др. Под ред. А.К. Соколова. М., 2004.
9. Керов В.В. Контент-анализ религиозно-этических комплексов как моделирование системы семантической сопряженности понятий: «деятельное страдание» в раннем староверии//Новые информационные ресурсы и технологии в исторических исследованиях и образовании. М., 2000.
10. Керов В.В. Отношение крайне правых к думским учреждениям предвоенного периода//Россия в XX веке: Люди, идеи, власть. М., 2002.
11. Ковальченко И. Д. Методы исторического исследования. М., 2003.
12. Количественные методы в исторических исследованиях. М., 1987.
13. Лотман Ю. М. Беседы о русской культуре. Быт и традиции русского дворянства (XVIII – начало XIX в.). СПб., 1994.
14. Лотман Ю. М. Статьи по семиотике культуры и искусства. СПб., 2002.
15. Лотман Ю.М. Символ в системе культуры // Символ в системе культуры. Труды по знаковым системам. Тарту, 1987. Вып. XXI.
16. Методы количественного анализа текстов нарративных источников. М., 1983.
17. Медушевская О. М. Методология когнитивной истории. М., 2008.
18. Плюханова М. Б. Сюжеты и символы Московского царства. СПб., 1995.
19. Успенский Б. А. История и семиотика (Восприятие времени как семиотическая проблема); Восприятие истории в Древней Руси и доктрина «Москва Третий Рим»; Царь и самозванец: самозванчество в России как культурно-исторический феномен; и др. // Успенский Б. А. Избранные труды. М., 1996. Т. 1.
20. Филюшкин А.И. Методологические инновации в современной российской исторической науке//ACTIO NOVA, 2000.
21. Elson D. K., Dames N., McKeown K. R. Extracting social networks from literary fiction // Proceedings of the 48th annual meeting of the association for computational linguistics. — Association for Computational Linguistics. 2010. — С. 138—147. Jockers M. L.,
22. Mimno D. Significant themes in 19th-century literature // Poetics. — 2013. — Т. 41, № 6. — С. 750—769. Koppel M., Argamon S., Shimoni A. R. Automatically categorizing written texts by author gender // Literary and Linguistic Computing. — 2002. — Т. 17, № 4. — С. 401—412.
23. Narrative framing of consumer sentiment in online restaurant reviews / D. Jurafsky [и др.] // First Monday. — 2014. — Т. 19, № 4.
24. Программа построения частотных словарей. http://alingva.ru/index.php/lingvosoft/12-ngramfrequency
25. mystem. Морфологический анализатор для русского языка. http://company.yandex.ru/technologies/mystem/
26. LSA. Латентно-семантический анализ текстовых данных. http://alingva.ru/index.php/lingvosoft/17-lsa
27. Tomita-пасрер. Инструмент для извлечения структурированных данных из текста на естественном языке. http://api.yandex.ru/tomita/
28. Модуль Perl Text::NSP. N-gram statistics and association measures. http://search.cpan.org/dist/TextNSP/lib/Text/NSP/Measures.pm
29. Mallet: MAchine Learning for LanguagE Toolkit. http://mallet.cs.umass.edu/
30. http://ruscorpora.ru — Национальный корпус русского языка.

7. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы

Требования к аудиториям (помещениям) для проведения занятий.

Учебные аудитории с компьютерным и проекционным оборудованием для демонстрации презентаций и выполнения индивидуальных заданий.

Требования к программному обеспечению общего пользования.
Пакет программ Microsoft Office 2010 Professional (Word, Excel, Access, PowerPoint), Google Chrome, Stata, SPSS Statistics 21, а также устойчивый источник Интернета для пользования онлайн-сервисами и тематическими сайтами.