АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Компьютерные методы анализа текста

Автор:
к.и.н., доцент кафедры истории экономики Кончаков Р.Б.

Код и наименование направления подготовки, профиля:

50.03.01 Искусства и гуманитарные науки
«Филология и перевод» (Liberal Arts)

Квалификация (степень) выпускника: бакалавр

Форма обучения: очная

Цель освоения дисциплины: сформировать способность использовать современное программное обеспечение и электронные ресурсы в планировании и реализации гуманитарных цифровых проектов

План курса:

№ 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.
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

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)
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:
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
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.

№ п/п	Наименование тем (разделов)	Содержание тем (разделов)
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.

Формы текущего контроля и промежуточной аттестации:

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

Промежуточная аттестация:
Зачет проводится в форме устного ответа на вопросы билета.

Основная литература:

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