МИНОБРНАУКИ РОССИИ

ФГБОУ ВО «УФИМСКИЙ ГОСУДАРСТВЕННЫЙ АВИАЦИОННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

ФАКУЛЬТЕТ ИНФОРМАТИКИ И РОБОТОТЕХНИКИ

Кафедра технической кибернетики

УТВЕРЖДАЮ:
Зав. кафедрой технической кибернетики

_____О.Я. Бежаева
(подпись, инициалы, фамилия)
«01» сентября 2022 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Интеллектуальный анализ текстовой информации
(наименование дисциплины)

47.04.01 Философия
шифр и наименование направления подготовки (специальности)

направленность (профиль) подготовки «Философия искусственного интеллекта»
наименование направленности (профиля, специализации)
квалификация: магистр

форма обучения: очно-заочная

***СОГЛАСОВАНО**: руководитель образовательной программы*

д.филос., наук, профессор БашГУ Елхова О.И.

Уфа – 2022

1.ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью освоения дисциплины «Интеллектуальный анализ текстовой информации» является ознакомление слушателей с методами обработки текста на естественном языке, а также методами обработки слабоструктурированных данных и извлечения информации. Предполагается знакомство с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов.

2.МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Настоящая дисциплина относится к вариативной части цикла дисциплин магистерской программы.

Изучение данной дисциплины базируется на следующих дисциплинах:

- ✓ Современные методы анализа данных,
- ✓ Современные методы принятия решений.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- ✓ Системный анализ и разработка сложных информационных систем;
- ✓ при выполнении проектов, подготовке ВКР

3. ТЕМАТИЧЕСКИЙ ПЛАН УЧЕБНОЙ ДИСЦИПЛИНЫ

Курс рассчитан на 32 часа аудиторной нагрузки, из них 12 часов лекций, 12 часов лабораторных работ 8 часов практических занятий, общим объемом 2 зачетные единицы (72 часа).

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Лабораторные работы	Практические занятия	
1	Введение в обработку естественного языка	9	1		2	6
2	Классификация и кластеризация текстов	11	1		2	8
3	Информационный	11	1		2	8

	поиск					
4	Введение в машинный перевод	11	1		2	8
5	Введение в извлечение информации	10	2	4		4
6	Методы машинного обучения в задаче извлечения информации	10	2	4		4
7	Извлечение мнений	10	2	4		4
Итого:	72	72	12	12	8	40

4.СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

1. Введение в обработку естественного языка.

Этапы анализа текста. Обзор основных приложений автоматического анализа текста (АОТ) (машинный перевод, информационный поиск, и т.д.). Слова, фразы, предложения, корпусы. Языковые модели. Автоматический морфологический анализ и синтез. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка

2. Классификация и кластеризация текстов.

Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы. Интеллектуальный анализ данных

3. Информационный поиск.

Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций

4. Введение в машинный перевод.

Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика.

5. Введение в извлечение информации.

Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов. Семантический анализ текста на основе семантико-синтаксических моделей управления. Разметка частей речи. Выделение именованных сущностей. Извлечение информации и отношений из текста. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов. Понятие лингвистического шаблона для извлечения информации. Инструментальные программные средства для построения систем извлечения информации из текстов. Извлечение знаний под управлением онтологий в системах класса OntosMiner.

6. Методы машинного обучения в задаче извлечения информации.

Формальные методы определения автора текста. Лингвостатистические параметры. Статистические методы атрибуции. Авторский инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов. Методы обнаружения спама: вероятностные и статистические, байесовский классификатор

7. Извлечение мнений.

Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению. Анализ тональности как задача классификации

Примеры заданий контрольной работы.

1. Составить регулярное выражение, удовлетворяющее заданным требованиям.
2. Построить наиболее вероятную цепочку тегов (скрытых состояний) в заданной скрытой марковской модели по указанному предложению.
3. Вывести формулу для коэффициентов заданного алгоритма сглаживания n-граммной языковой модели.

4. Построить символьную триграммную языковую модель по заданному корпусу и с ее помощью построить распознаватель языка документа.

5. Вычислить перплексию n-граммной языковой модели с заданным сглаживанием.

6. На основе заданной обучающей выборки построить марковскую модель максимальной энтропии для выделения заданных именованных сущностей (имен собственных, географических названий и т. д.) из текста.

5.ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Основными образовательными технологиями являются: работа в группах на лабораторных и практических занятиях, проектный метод.

6.УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1 Основная литература

1. Aggarwal C. C., Zhai C. X. (ed.). Mining text data. – Springer Science & Business Media, 2012. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=54151

2. Ceri S. et al. Web Information Retrieval. Springer, 2013. 287 p. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=77020

6.2 Дополнительная литература

1. Munzert S., Rubba C., MeiЯner P., Nyhuis D. Automated data collection with R: A practical guide to web scraping and text mining. – John Wiley & Sons, 2014. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=72676

2. Goker A., Davies J. (ed.). Information retrieval: searching in the 21st century. – John Wiley & Sons, 2009. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=33746

6.3 Ресурсы информационно-телекоммуникационной сети «Интернет»

Для извлечения информации используются следующие сайты:

wikipedia.org – онлайн энциклопедия

twitter.com – сервис блогов

vk.com – социальная сеть с богатым API для доступа к информации

www.tripadvisor.ru – сайт отзывов

7.РЕКОМЕНДАЦИИ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Самостоятельная работа может рассматриваться как организационная форма обучения – система педагогических условий, обеспечивающих управление учебной деятельностью по освоению знаний и умений в области учебной деятельности без посторонней помощи. Студенту нужно четко понимать, что самостоятельная работа – не просто обязательное, а необходимое условие для получения знаний по дисциплине и развитию компетенций, необходимых в будущей профессиональной деятельности.

Самостоятельная работа проводится с целью:

− систематизации и закрепления полученных на лекциях теоретических знаний;

− углубления и расширения теоретических знаний;

− формирования умений использовать нормативную, правовую, справочную документацию и специальную литературу;

− развития познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности;

− формирования самостоятельности мышления, способностей к саморазвитию, самосовершенствованию и самореализации;

− формирования практических (общеучебных и профессиональных) умений и навыков;

− развития исследовательских умений;

− получения навыков эффективной самостоятельной профессиональной (практической и научно-теоретической) деятельности.

8.МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

и информационные технологии, используемые при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения информационных справочных систем (при необходимости).

Практические занятия проводятся в компьютерных классах. На лекциях и практических занятиях используется проектор. Для успешного освоения дисциплины, студент использует следующие программные средства: Python (пакеты scipy и numpy, сборка Anaconda, Pandas, Scikit-learn и др.), инструменты R и RStudio.

9. ОСОБЕННОСТИ ОРГАНИЗАЦИИ ОБУЧЕНИЯ ДЛЯ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) могут предлагаться следующих варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

1) для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат);

индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

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.

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