Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Курганский государственный университет»

Кафедра «Программное обеспечение автоматизированных систем»

УТВЕРЖДАЮ:
Ректор
_____________ Н.В. Дубив
«31» августа 2020 г.

Рабочая программа учебной дисциплины

ОБРАБОТКА И АНАЛИЗ ТЕКСТОВ

образовательной программы высшего образования –
программы магистратуры
09.04.04 Программная инженерия

направленность
Методы и алгоритмы интеллектуальной обработки данных
в информационно-вычислительных системах

формы обучения – очная, заочная

Курган 2020

Рабочая программа дисциплины «Обработка и анализ текстов» составлена в соответствии с учебным планом по программе магистратуры «Программная инженерия» (Методы и алгоритмы интеллектуальной обработки данных в информационно-вычислительных системах), утвержденным для очной и заочной форм обучения «30» августа 2020 года.

Рабочая программа дисциплины одобрена на заседании кафедры «Программное обеспечение автоматизированных систем» «30» августа 2020 года, протокол № 1.

Рабочую программу составила:

Доцент кафедры
«Программное обеспечение
автоматизированных систем», к.т.н. Н.В. Агапова

Согласовано:

Заведующий кафедрой
«Программное обеспечение
автоматизированных систем» Т.Р. Змызгова

Начальник Управления образовательной
деятельности С.Н. Синицын

Специалист по учебно-методической
работе Учебно-методического отдела Г.В. Казанкова

1. ОБЪЕМ ДИСЦИПЛИНЫ

Всего: 5 зачетных единиц трудоемкости (180 академических часов)

Очная форма обучения

Вид учебной работы	На всю дисциплину	Семестр
		3
Аудиторные занятия (контактная работа с преподавателем), всего часов в том числе:	**72**	**72**
Лекции	24	24
Лабораторные работы	48	48
Аудиторные занятия в интерактивной форме, часов	-	-
Самостоятельная работа, всего часов в том числе:	**108**	**108**
Подготовка к экзамену	27	27
Другие виды самостоятельной работы	81	81
Вид промежуточной аттестации	**экзамен**	**экзамен**
Общая трудоемкость дисциплины и трудоемкость по семестрам, часов	**180**	**180**

Заочная форма обучения

Вид учебной работы	На всю дисциплину	Семестр
		3
Аудиторные занятия (контактная работа с преподавателем), всего часов в том числе:	**20**	**20**
Лекции	10	10
Лабораторные работы	10	10
Аудиторные занятия в интерактивной форме, часов	-	-
Самостоятельная работа, всего часов в том числе:	**160**	**160**
Контрольная работа	18	18
Подготовка к экзамену	27	27
Другие виды самостоятельной работы	115	115
Вид промежуточной аттестации	**экзамен**	**экзамен**
Общая трудоемкость дисциплины и трудоемкость по семестрам, часов	**180**	**180**

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина «Обработка и анализ текстов» относится к части блока 1, формируемой участниками образовательных отношений, модулю «Прикладные задачи интеллектуального анализа данных (элективный модуль)», дисциплина по выбору.

Программа составлена с учетом межпредметных связей с учебными дисциплинами. Основой для изучения учебной дисциплины являются следующие учебные дисциплины «Структуры и алгоритмы обработки данных», «Интеллектуальный анализ данных», «Управление данными», «Архитектуры информационно-вычислительных систем», «Методологии и технологии информационно-вычислительных систем».

Учебная дисциплина «Обработка и анализ текстов» знакомит студентов с методами хранения, сбора, обработки, выдачи больших данных.

Результаты обучения по дисциплине необходимы для изучения дисциплин: «Научно-исследовательская работа (производственная практика)» и выполнения выпускной квалификационной работы.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Целью освоения дисциплины «Обработка и анализ текстов» 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.

Компетенции, формируемые в результате освоения дисциплины:

- Владение методами планирования и обработки результатов экспериментальных исследований (ПК-2);

- Способность разрабатывать и использовать программное обеспечение для моделирования, анализа, распознавания и обработки информации, в том числе - в системах искусственного интеллекта (ПК-3)

В результате изучения дисциплины обучающийся должен:

Знать: математические методы обработки данных (для ПК-2, ПК-3).

Уметь: выполнять исследования процессов создания, накопления и обработки информации, включая анализ и создание моделей данных и знаний, языков их описания и манипулирования (для ПК-2, ПК-3).

Владеть: новыми методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности (для ПК-2, ПК-3).

4. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1. Учебно-тематический план

Очная форма обучения

№	Наименование раздела	Количество часов контактной работы с преподавателем	
		Лекции	Лабораторные занятия
1	Введение в обработку естественного языка	2	8
2	Классификация и кластеризация текстов	2	8
3	Информационный поиск	4	4
4	Введение в машинный перевод	4	6
5	Введение в извлечение информации	4	4
6	Методы машинного обучения в задаче извлечения информации	4	12
7	Извлечение мнений	4	6
	Всего:	**24**	**48**

Заочная форма обучения

№	Наименование раздела	Количество часов контактной работы с преподавателем	
		Лекции	Лабораторные занятия
1	Введение в обработку естественного языка	2	2
2	Классификация и кластеризация текстов	2	2
3	Информационный поиск	2	2
4	Введение в машинный перевод	1	1
5	Введение в извлечение информации	1	1
6	Методы машинного обучения в задаче извлечения информации	1	1
7	Извлечение мнений	1	1
	Всего:	**10**	**10**

4.2. Содержание лекционных занятий

Тема 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

Тема 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

Тема 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

Тема 4. **Введение в машинный перевод**

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

Тема 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.

Тема 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

***Тема 7.* Извлечение мнений**

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

Вычислительные подходы к моделированию.

4.3. Лабораторные работы

Номер раздела, темы	Наименование раздела	Наименование лабораторной работы	Норматив времени, час.	
			Очная форма обучения	Заочная форма обучения
1	Введение в обработку естественного языка	1.Разработка автоматизированной системы формирования словаря естественного языка 2. Синтаксический анализ текстов естественного языка	8	2
2	Классификация и кластеризация текстов	3. Классификация текстов писателей 4. Сегментация текста на примере базы данных договоров	8	2
3	Информационный поиск	5. Поиск похожих объектов	4	2
		Рубежный контроль 1	2	-
4	Введение в машинный перевод	6. Создание переводчиков с английского, на английский	4	1
5	Введение в извлечение информации	7. Онлайновые алгоритмы. Сопоставление предложений с поисковыми запросами	4	1
6	Методы машинного обучения в задаче извлечения информации	8. Семантико-синтаксический анализ текстов естественного языка 9. Диалоговая система с поддержкой естественного языка – генерация текста	12	1
7	Извлечение мнений	10. Сравнение методов обучения	4	1
	Рубежный контроль 2		2	-
		Всего:	**48**	**10**

4.5. Контрольная работа

(для обучающихся заочной формы обучения)

Контрольная работа посвящена решению задач методами обработки текста на естественном языке, а также методами обработки слабоструктурированных данных и извлечения информации. Часть заданий посвящена знакомству с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов. Контрольная работа выполняется обучающимися по вариантам заданий или по теме, предложенной обучающимся и согласованной с преподавателем.

1. Распознавание рукописной цифры
2. Классификация текстов писателей
3. Классификатор спама
4. Оценка стоимости квартир
5. Создание переводчиков с английского, на английский
6. Сеть, отвечающая на вопросы
7. Генерация текста - чатбот
8. Сегментация (токенизация) текста
9. Распознавание речи
10. Синтез речи

Основная учебная цель: закрепление теоретических знаний, полученных в процессе изучения дисциплины и приобретение практических навыков по решению задач на обработку текста.

Требования к содержанию контрольной работы

Контрольная работа должна содержать визуальное приложение и комплект документации:

пояснительная записка;

файл с реализованной программой.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Во время лекций по дисциплине студентам рекомендуется конспектировать теоретический материал, отмечая важные моменты, на которые заострил внимание преподаватель, участвовать в опросах и дискуссиях. Перед лекцией необходимо повторить выданный материал, зафиксировать непонятные места, чтобы обсудить их на занятии. Конспект лекций представлен в виде мультимедийных презентаций и включен в состав методического комплекса дисциплины.

Лабораторный практикум включает практические задания по основным одиннадцати разделам дисциплины. Все работы выполняются в соответствии с заданием, выданным преподавателем.

Преподавателем запланировано применение на лабораторных занятиях технологий развивающейся кооперации, коллективного взаимодействия, разбора конкретных ситуаций. Поэтому приветствуется групповой метод выполнения лабораторных работ и защиты отчетов, а также взаимооценка и обсуждение результатов выполнения лабораторных работ.

Для текущего контроля успеваемости по очной форме обучения преподавателем используется балльно-рейтинговая система контроля и оценки академической активности.

Выполнение самостоятельной работы подразумевает самостоятельное изучение разделов дисциплины, подготовку к лабораторным занятиям, к рубежным контролям (для обучающихся очной формы обучения), выполнение контрольной работы (для обучающихся заочной формы обучения), подготовку к экзамену.

Рекомендуемый режим самостоятельной работы

Наименование вида самостоятельной работы	Рекомендуемая трудоемкость, акад. час.	
	Очная форма обучения	Заочная форма обучения
Самостоятельное изучение тем дисциплины:	**53**	95
Компьютерная лингвистика: задачи, подходы Лингвистические ресурсы: построение и применение Сложности моделирования естественного языка Подходы к построению модулей и систем компьютерной лингвистики	9	16
Морфологический анализ текста Обзор модулей морфологического анализа. Методы хранения словарей Анализ несловарных слов Особенности ононимии, её разрешение	9	17
Извлечение информации из текстов Специфика задач, подходы к решению Извлечение атрибутов, текстов и фактов Инструментальные системы для извлечения информации Извлечение терминологической информации	9	16
Автоматические методы извлечения тональности Сложности анализа тональности текстов Анализ документов Анализ по аспектам Тестирование систем анализа тональности текстов	9	10
Обзор вероятностных тематических моделей Основы тематического моделирования Интерпретируемость тем Определение числа тем Модальности, зависимости Связи между документами Иерархии тем Совстречаемость слов Разведочный информационный поиск	9	10
Методы машинного обучения в задаче извлечения информации Лингвостатистические параметры Авторский инвариант, лингвистические спектры Методы обнаружения спама	4	10
Извлечение мнений Извлечение мнений: особенности и подходы к решению	4	16
Подготовка к контрольной работе	-	**18**
Подготовка к лабораторным работам (по 2 ч. на каждое занятие для ОФО и по 4 ч. на каждое занятие для ЗФО)	**24**	**20**
Подготовка к рубежному контролю 1	**2**	-
Подготовка к рубежному контролю 2	**2**	-
Подготовка к экзамену	**27**	**27**
Всего:	**108**	**160**

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

6.1. Перечень оценочных средств

1. Балльно-рейтинговая система контроля и оценки академической активности обучающихся в КГУ (для очной формы обучения)
2. Отчеты обучающихся по лабораторным занятиям
3. Тестовые задания
4. Банк заданий к рубежным контролям № 1, № 2 (для очной формы обучения).
5. Контрольная работа (для заочной формы обучения)
6. Вопросы к экзамену

6.2. Система балльно-рейтинговой оценки работы обучающихся по дисциплине

Очная форма обучения

<table>
<tr><th>№</th><th>Наименование</th><th colspan="6">Содержание</th></tr>
<tr><td rowspan="4">1</td><td rowspan="4">Распределение баллов за семестры по видам учебной работы, сроки сдачи учебной работы (доводятся до сведения обучающихся на первом учебном занятии)</td><td colspan="6">Распределение баллов,
3 семестр</td></tr>
<tr><td>Вид учебной работы:</td><td>Посещение лекций</td><td>Выполнение и защита результатов лабораторных работ</td><td>Рубежный контроль №1</td><td>Рубежный контроль №2</td><td>Экзамен</td></tr>
<tr><td>Балльная оценка:</td><td>До 12</td><td>10 х 4 б = 40 б</td><td>9</td><td>9</td><td>30</td></tr>
<tr><td>Примечания:</td><td>12 лекций по 1 баллу</td><td>10 занятий по 4 балла</td><td></td><td></td><td></td></tr>
<tr><td>2</td><td colspan="3">Критерий пересчета баллов в традиционную оценку по итогам работы в семестре и экзамена</td><td colspan="4">60 и менее баллов – неудовлетворительно;
61…73 – удовлетворительно;
74… 90 – хорошо;
91…100 – отлично.</td></tr>
</table>

3	Критерии допуска к промежуточной аттестации, возможности получения автоматического зачета (экзаменационной оценки) по дисциплине, возможность получения бонусных баллов	Для допуска к промежуточной аттестации (экзамену) обучающийся должен набрать не менее 50 баллов, выполнить все лабораторные работы и выполнить контрольную работу для ЗФО. Для получения «автоматически» экзаменационной оценки «удовлетворительно» обучающемуся необходимо набрать 68 баллов. По согласованию с преподавателем обучающемуся, набравшему минимум 68 баллов, могут быть добавлены дополнительные (бонусные) баллы за активность на консультациях, активное участие в научной и методической работе, оригинальность принятых решений в ходе выполнения лабораторных работ, за участие в значимых учебных и внеучебных мероприятиях кафедры и выставлена за экзамен «автоматически» оценка «хорошо» или «отлично».
4	Формы и виды учебной работы для неуспевающих (восстановившихся на курсе обучения) обучающихся для получения недостающих баллов в конце семестра	В случае если к промежуточной аттестации (экзамену) набрана сумма менее 50 баллов, обучающемуся необходимо набрать недостающее количество баллов за счет выполнения дополнительных заданий, до конца последней (зачетной) недели семестра. При этом необходимо проработать материал всех пропущенных лабораторных работ. Формы дополнительных заданий (назначаются экзаменатором): - выполнение и защита пропущенной лабораторной работы (при невозможности дополнительного проведения лабораторной работы преподаватель устанавливает форму дополнительного задания по тематике пропущенной лабораторной работы самостоятельно) – до 4 баллов. Ликвидация академических задолженностей, возникших из-за разности в учебных планах при переводе или восстановлении, проводится путем выполнения дополнительных заданий, форма и объем которых определяется преподавателем.

6.3. Процедура оценивания результатов освоения дисциплины

Рубежные контроли и экзамен проводятся в форме беседы по вопросам.

Перед проведением каждого рубежного контроля преподаватель прорабатывает с обучающимися основной материал соответствующих разделов дисциплины в форме краткой лекции-дискуссии.

При проведении рубежного контроля обучающемуся предлагается из списка один теоретический вопрос и одно практическое задание или два теоретических вопроса. На подготовку к ответу обучающемуся отводится время не менее 40 минут. Преподаватель оценивает в баллах ответ каждого обучающегося по количеству правильных ответов и заносит в ведомость учета текущей успеваемости.

Билеты на экзамен состоят из 2 вопросов и практического задания. Ответы на каждый вопрос оцениваются до 10 баллов, выполнение практического задания оценивается до 10 баллов. Время, отводимое обучающемуся на подготовку к ответу на экзаменационный билет, составляет 1 астрономический час.

Результаты текущего контроля успеваемости и экзамена заносятся преподавателем в зачетно-экзаменационную ведомость, которые сдаются в организационный отдел института в день экзамена, а также выставляются в зачетную книжку обучающегося.

6.4. Примеры оценочных средств для рубежных контролей и экзамена

6.4.1 Примеры заданий для рубежного контроля №1

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Проверка статистических гипотез для поиска словосочетаний. Проверка по критерию Стьюдента.
6. Проверка статистических гипотез для поиска словосочетаний. Критерий согласия Пирсона
7. Проверка статистических гипотез для поиска словосочетаний. Отношение правдоподобия
8. Проверка статистических гипотез для поиска словосочетаний. Информационно-теоретический подход к поиску словосочетаний
9. Модель N-грамм. Оценка вероятности высказывания
10. Модель N-грамм. Сглаживание (Лапласа и Откат)
11. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
12. Задача определения частей речи. Существующие подходы
13. Использование скрытой марковской модели для определения частей речи
14. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм
15. Скрытые марковские модели. Наиболее правдоподобное объяснение. Алгоритм Витерби
16. Модели классификации. Наивный байесовский классификатор
17. Модели классификации. Логистическая регрессия
18. Модели классификации. Модель максимальной энтропии

19. Модели классификации. Марковская модель максимальной энтропии
20. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика
21. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.

6.4.2 Примеры заданий для рубежного контроля №2

22. Синтаксический разбор. Разбор сверху вниз и снизу вверх
23. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
24. Фрагментирование
25. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
26. Моделирование языка. Обучение стохастических КС грамматик
27. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
28. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
29. Лексическая семантика. WordNet. Значения слов
30. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка.
31. Разрешение лексической многозначности. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
32. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
33. Семантическая близость слов. Подходы на основе статистик. Методы оценки качества
34. Вопросно-ответные системы. Общая архитектура. Обработка запроса
35. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
36. Вопросно-ответные системы. Общая архитектура. Обработка ответа.
37. Автоматическое реферирование. Общая архитектура
38. Машинный перевод. Классические подходы
39. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз (если слова выровнены). Декодирование
40. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
41. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
42. Статистический машинный перевод. Методы оценки качества. BLUE.

6.4.3 Примерный перечень вопросов для экзамена

1. Этапы анализа текста.
2. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ.
3. Морфологические процессоры для русского языка
4. Классификация текстов
5. Обзор методов машинной классификации.
6. Особенности кластеризации текстов.
7. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы.
8. Индексирование текстов для информационного поиска.
9. Векторная модель документа.
10. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин.
11. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска.
12. Основные стратегии сжатия текста.
13. Обзорное реферирование. Оценка качества аннотаций
14. Стратегии машинного перевода, основанного на лингвистических правилах.
15. Статистический машинный перевод: особенности и виды.
16. Принципы создания статистического переводчика.
17. Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов.
18. Семантический анализ текста на основе семантико-синтаксических моделей управления.
19. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов.
20. Лингвистический шаблон для извлечения информации.
21. Инструментальные программные средства для построения систем извлечения информации из текстов.
22. Извлечение знаний под управлением онтологий в системах класса OntosMiner.
23. Формальные методы определения автора текста. Лингвостатистические параметры.
24. Статистические методы атрибуции.
25. Применение методов кластеризации и классификации для установления авторства текстов.
26. Методы обнаружения спама: вероятностные и статистические, байесовский классификатор
27. Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению.

28. Анализ тональности как задача классификации
29. Вычислительные подходы к моделированию.

6.4.4 Примеры практических заданий экзамена

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.

6.5. Фонд оценочных средств

Полный банк заданий для текущего, рубежных контролей и промежуточной аттестации по дисциплине, показатели, критерии, шкалы оценивания компетенций, методические материалы, определяющие процедуры оценивания образовательных результатов, приведены в учебно-методическом комплексе дисциплины.

7. ОСНОВНАЯ И ДОПОЛНИТЕЛЬНАЯ УЧЕБНАЯ ЛИТЕРАТУРА

7.1. Основная учебная литература

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. Пособие / Большакова, Е. И. Воронцов, К.В. и др. – М.: Изд-во НИУ ВШЭ. 2017. – 269 с.
https://www.hse.ru/data/2017/08/12/1174382135/NLP_and_DA.pdf
2. Крапивин, Ю.Б. Естественно языковой интерфейс интеллектуальных систем. Лабораторный практикум: пособие / Крапивин, Ю.Б. – Минск: БГУИР, 2020. – 64 с. ISBN 978-985-543-680-6
https://libeldoc.bsuir.by/bitstream/123456789/50730/1/Krapivin_Estestvenno_yazikovoi.pdf

7.2. Дополнительная учебная литература

Информационные ресурсы: Журналы по обработке текстовой информации (ComputationalLinguistics, ACL Journal), труды конференций (ACL, EACL, COLING, EMNLP, Диалог), доступные через Internet, электронные конспекты лекций, разработанные для данного курса.

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

8. РЕСУРСЫ СЕТИ «ИНТЕРНЕТ», НЕОБХОДИМЫЕ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1. Электронная библиотека КГУ http://dspace.kgsu.ru/xmlui/

9. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ И ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ

1. ЭБС «Лань»
2. ЭБС «Консультант студента»
3. ЭБС «Znanium.com»
4. «Гарант» - справочно-правовая система

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Материально - техническое обеспечение по реализации дисциплины осуществляется с требованиями ФГОС ВО по данной образовательной программе.

11. ДЛЯ ОБУЧАЮЩИХСЯ С ИСПОЛЬЗОВАНИЕМ ДИСТАНЦИОННЫХ ОБРАЗОВАТЕЛЬНЫХ ТЕХНОЛОГИЙ

При использовании электронного обучения и дистанционных образовательных технологий (далее ЭО и ДОТ) занятия полностью или частично проводятся в режиме онлайн. Объем дисциплины и распределение нагрузки по видам работ соответствует п. 4.1. Распределение баллов соответствует п. 6.2 либо может быть изменено в соответствии с решением кафедры, в случае перехода на ЭО и ДОТ в процессе обучения. Решение кафедры об используемых технологиях и системе оценивания достижений обучающихся принимается с учетом мнения ведущего преподавателя и доводится до обучающихся.

Аннотация
рабочей программы учебной дисциплины

ОБРАБОТКА И АНАЛИЗ ТЕКСТОВ
образовательной программы высшего образования –
программы магистратуры

09.04.04 Программная инженерия
Направленность:
Методы и алгоритмы интеллектуальной обработки данных в информационно-вычислительных системах

Форма обучения: очная, заочная

Трудоемкость освоения дисциплины – 5 зач. ед. (144 акад. часа)
Семестры: 3-й (очная, заочная форма обучения)
Промежуточная аттестация: экзамен.

Содержание дисциплины

Целью освоения дисциплины «Обработка и анализ текстов» 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.