МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное
учреждение высшего образования
«Мурманский арктический государственный университет»
(ФГБОУ ВО «МАГУ»)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.01.05 Анализ текстовых данных

(название дисциплины (модуля) в соответствии с учебным планом)

основной профессиональной образовательной программы
по направлению подготовки

01.03.02 Прикладная математика и информатика
направленность Управление данными и машинное обучение

(код и наименование направления подготовки
с указанием направленности (наименования магистерской программы))

высшее образование – бакалавриат

уровень профессионального образования: высшее образование – бакалавриат / высшее образование – специалитет, магистратура / высшее образование – подготовка кадров высшей квалификации

бакалавр

квалификация

очная

форма обучения

2021

год набора

Составитель(и):

Лазарева Ирина Михайловна,
доцент, к.ф.-м.н.,
зав. кафедрой МФиИТ

Утверждено на заседании кафедры
математики, физики и информационных
технологий факультета
математических и естественных наук
(протокол № 07 от 12.04.2021)

Переутверждено на заседании кафедры
математики, физики и информационных
технологий факультета
математических и естественных наук
(протокол № 09 от 02.07.2021)

Зав. кафедрой _____ Лазарева И.М.
подпись Ф.И.О.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ).

Цель – формирование понимания современных подходов к анализу текстовых данных и обработке текстов методами машинного обучения.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

В результате освоения дисциплины (модуля) формируются следующие компетенции:

ПК-1: Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим прикладным исследованиям

ПК-2: Способен работать в составе научно-исследовательского и производственного коллектива и решать задачи профессиональной деятельности

ПК-3: Способен к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения

Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ПК-1: Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим прикладным исследованиям **ПК-2:** Способен работать в составе научно-исследовательского и производственного коллектива и решать задачи профессиональной деятельности **ПК-3:** Способен к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения	ПК-1.1 Понимает содержательную постановку задачи ПК-1.2 Умеет грамотно отбирать значимые данные ПК-1.3 Умеет представлять результаты своей деятельности с учетом уровня аудитории ПК-2.1 Формулирует задачи в рамках проекта и определяет ожидаемые результаты ПК-2.2 Обеспечивает модульность выполнения задачи с учетом имеющихся ресурсов ПК-2.3 Обеспечивает пользовательскую привлекательность создаваемого программного продукта ПК-3.1 Разрабатывает алгоритм решения поставленной задачи выбранным методом ПК-3.2 Выбирает и обосновывает выбор языковой среды ПК-3.3 Использует современную языковую среду для реализации сложных алгоритмов ПК-3.4 Решает задачу тестирования программного продукта	*Знать:* – основные способы получения и обработки информации, необходимой для профессиональной деятельности; – основные парадигмы машинного обучения; – модели и методы предобработки текстовых данных; – методы оценки качества моделей машинного обучения *Уметь:* – применять методы машинного обучения для решения задач анализа текстовых данных; – оценивать качество моделей машинного обучения; – обрабатывать и анализировать результаты эксперимента, проводить расчеты по экспериментальным данным с использованием компьютерных программ *Владеть:* – навыком исследования и моделирования предметной области; – владеть терминологией машинного обучения и искусственных нейронных сетей; – владеть инструментальными средствами для построения моделей машинного обучения с учителем; – навыками работы с наиболее распространенными прикладными пакетами для обработки текстовых данных; – основными методами, способами и средствами получения, хранения, переработки информации

3. УКАЗАНИЕ МЕСТА ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ.

Дисциплина «Анализ текстовых данных» относится к части, формируемой участниками образовательных отношений (Б1.В.01.05) по направлению подготовки 01.03.02 Прикладная математика и информатика, направленность (профиль) Управление данными и машинное обучение.

4. ОБЪЕМ ДИСЦИПЛИНЫ (МОДУЛЯ) В ЗАЧЕТНЫХ ЕДИНИЦАХ С УКАЗАНИЕМ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ИЛИ АСТРОНОМИЧЕСКИХ ЧАСОВ, ВЫДЕЛЕННЫХ НА КОНТАКТНУЮ РАБОТУ ОБУЧАЮЩИХСЯ С ПРЕПОДАВАТЕЛЕМ (ПО ВИДАМ УЧЕБНЫХ ЗАНЯТИЙ) И НА САМОСТОЯТЕЛЬНУЮ РАБОТУ ОБУЧАЮЩИХСЯ.

Общая трудоемкость дисциплины (модуля) составляет 3 зачетные единицы или 108 часа (из расчета 1 ЗЕ = 36 часов).

Курс	Семестр	Трудоемкость в ЗЕ	Общая трудоемкость (час)	Контактная работа			Всего контактных часов	Из них:		Кол-во часов на СРС		Кол-во часов на контроль	Форма контроля
				ЛК	ПР	ЛБ		В интерактивной форме*	В форме практической подготовки*	Общее количество часов на СРС	из них – на курсовую работу		
3	6	3	108	20	-	34	50	8	18	54	-	-	зачет
Итого:		**3**	**108**	**20**	**-**	**34**	**50**	**8**	**18**	**54**	**-**	**-**	**зачет**

Интерактивная форма реализуется в виде проблемных лекций и проектной деятельности по темам дисциплины.

Практическая подготовка реализуется в виде лабораторных работ.

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ), СТРУКТУРИРОВАННОЕ ПО ТЕМАМ (РАЗДЕЛАМ) С УКАЗАНИЕМ ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ИЛИ АСТРОНОМИЧЕСКИХ ЧАСОВ И ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ.

№ п/п	Наименование раздела, темы	Контактная работа			Всего контактных часов	Из них		Кол-во часов на СРС	Кол-во часов на контроль
		ЛК	ПР	ЛБ		в интерактивной форме	В форме практической подготовки		
Раздел 1. Введение в анализ текстовых данных									
1.	Введение в анализ текстов, базовые методы предобработки и выделения признаков	2	-	4	6	-	2	10	
2.	Неглубокие векторные представления слов	4	-	4	8	2	2	10	
3.	Классификация текстов	2	-	4	6	-	2		

№ п/п		Контактная работа			контактну	Из них		во часов	в часов
Раздел 2. Задачи текстового анализа									
4.	Разметка последовательности	4	-	6	10	2	4	12	
5.	Машинный перевод	4	-	8	12	2	4	12	
6.	Тематическое моделирование	4	-	8	12	2	4	12	
	Зачет								-
	ИТОГО:	**20**	**-**	**34**	**54**	**8**	**18**	**54**	**-**

Содержание дисциплины (модуля)

Тема 1. Введение в анализ текстов, базовые методы предобработки и выделения признаков.

Задачи анализа текстов, специфика, история. Базовая предобработка текстов. Простейшие текстовые признаки: “мешок слов”, TF-IDF. Предобработка текстовых данных: регулярные выражения.

Тема 2. Неглубокие векторные представления слов.

Идея векторных представлений, one-hot-векторы, SVD. Модель word2vec, методы её обучения. Оптимизации обучения word2vec: SGNS и иерархический softmax. Модель GloVe. Модель и библиотека FastText, приём Hashing Trick.

Тема 3. Классификация текстов.

Задача классификация текстов. Логистическая регрессия на счётчиках и TF-IDF. Неглубокие векторные представления документов. Библиотека FastText для классификации текстов. CNN для классификации текстов. Работа с обучающими данными.

Тема 4. Разметка последовательности.

Счетные языковые модели. Морфологический анализ, скрытые Марковские модели. Нейросетевые языковые модели. Рекуррентные нейронные сети (RNN). Генерация текстов. Извлечение именованных сущностей (NER). Перекрестное обучение.

Тема 5. Машинный перевод.

Модели класса кодировщик-декодировщик. Механизм внимания. Модель Трансформер. Метрики качества в машинном переводе. Улучшение качества машинного перевода.

Тема 6. Тематическое моделирование.

Постановка задачи тематического моделирования. Модель PLSA. EM-алгоритм. Модель ARTM. Модель LDA. Модель M-ARTM. Технические аспекты обучения TM.

6. ПЕРЕЧЕНЬ УЧЕБНО-МЕТОДИЧЕСКОГО ОБЕСПЕЧЕНИЯ, НЕОБХОДИМОГО ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ).

Основная литература:

1. Анализ данных : учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2018. — 490 с. — (Бакалавр. Академический курс). — ISBN 978-5-534-00616-2. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/412967
2. Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2018. — 174 с. — (Авторский учебник). — ISBN 978-5-9916-5009-0. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/413060
3. Чубукова, И.А. Data Mining : учебное пособие : [16+] / И.А. Чубукова. – 2-е изд., испр. – Москва : Интернет-Университет Информационных Технологий (ИНТУИТ) : Бином. Лаборатория знаний, 2008. – 383 с. – (Основы информационных технологий). – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=233055. – ISBN 978-5-94774-819-2. – Текст : электронный.

Дополнительная литература:

1. Каган, Е.С. Прикладной статистический анализ данных : учебное пособие : [16+] / Е.С. Каган ; Кемеровский государственный университет. – Кемерово : Кемеровский государственный университет, 2018. – 235 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=573550. – Библиогр.: с. 184-186. – ISBN 978-5-8353-2413-2. – Текст : электронный.
2. Келлехер, Д. Наука о данных: базовый курс : [16+] / Д. Келлехер, Б. Тирни ; науч. ред. З. Мамедьяров ; пер. с англ. М. Белоголовского. – Москва : Альпина Паблишер, 2020. – 224 с. : схем., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=598235. – ISBN 978-5-9614-3170-4. – Текст : электронный.
3. Парфенов, Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов. — М. : Издательство Юрайт, 2018. — 121 с. — (Серия : Университеты России). — ISBN 978-5-534-03408-0. — Режим доступа : www.biblio-online.ru/book/628DAC6C-ECBF-45B3-BD23-F6B57148D18F.

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ).

В образовательном процессе используются:

- учебные аудитории для проведения учебных занятий, оснащенные оборудованием и техническими средствами обучения: учебная мебель, ПК, оборудование для демонстрации презентаций, наглядные пособия;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду МАГУ.

7.1 ПЕРЕЧЕНЬ ЛИЦЕНЗИОННОГО И СВОБОДНО РАСПРОСТРАНЯЕМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ:

Лицензионное программное обеспечение:

- Операционная система: MS Windows версии 7 и выше;
- Программные средства, входящие в состав офисного пакета: MS Office (Word, Excel, Access, Publisher, PowerPoint);
- Программные обеспечение: MS OfficeVisio, MS ACCESS, MS SQL SERVER 2008, Visual Studio 2010.

Свободно распространяемое программное обеспечение:

- Программные обеспечение: MongoDB.
- Программы для просмотра документов: Abobe Acrobar Reader, DJVU Reader;
- Среда логического проектирования структуры базы данных Erwin;
- Браузер: Google Chrome;
- Архиватор: 7Zip.

7.2 ЭЛЕКТРОННО-БИБЛИОТЕЧНЫЕ СИСТЕМЫ:

- ЭБС «Издательство Лань» [Электронный ресурс]: электронная библиотечная система / ООО «Издательство Лань». – Режим доступа: https://e.lanbook.com/;
- ЭБС «Электронная библиотечная система ЮРАЙТ» [Электронный ресурс]: электронная библиотечная система / ООО «Электронное издательство ЮРАЙТ». – Режим доступа: https://biblio-online.ru/;
- ЭБС «Университетская библиотека онлайн» [Электронный ресурс]: электронно-периодическое издание; программный комплекс для организации онлайн-доступа к лицензионным материалам / ООО «НексМедиа». – Режим доступа: https://biblioclub.ru/

7.3 СОВРЕМЕННЫЕ ПРОФЕССИОНАЛЬНЫЕ БАЗЫ ДАННЫХ:

- Информационно-аналитическая система SCIENCE INDEX

- Электронная база данных Scopus
- Базы данных компании CLARIVATE ANALYTICS

7.4. ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ:
- Справочно-правовая информационная система Консультант Плюс http://www.consultant.ru/
- ООО «Современные медиа технологии в образовании и культуре». http://www.informio.ru/

8. ИНЫЕ СВЕДЕНИЯ И МАТЕРИАЛЫ НА УСМОТРЕНИЕ ВЕДУЩЕЙ КАФЕДРЫ.

1. http://www.machinelearning.ru/ - Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных

9. ОБЕСПЕЧЕНИЕ ОБРАЗОВАНИЯ ДЛЯ ЛИЦ С ОВЗ.

Для обеспечения образования инвалидов и лиц с ограниченными возможностями здоровья реализация дисциплины может осуществляться в адаптированном виде, с учетом специфики освоения и дидактических требований, исходя из индивидуальных возможностей и по личному заявлению обучающегося.