Федеральное государственное бюджетное образовательное учреждение высшего образования
«Алтайский государственный технический университет им. И.И. Ползунова»

СОГЛАСОВАНО

Декан ФИТ А.С. Авдеев

Рабочая программа дисциплины

Код и наименование дисциплины: **Б1.В.ДВ.1.1 «Методы моделирования естественных языков»**

Код и наименование направления подготовки (специальности): **09.06.01 Информатика и вычислительная техника**

Направленность (профиль, специализация): **Математическое моделирование, численные методы и комплексы программ**

Статус дисциплины: **дисциплины (модули) по выбору**

Форма обучения: **очная**

Статус	Должность	И.О. Фамилия
Разработал	профессор	Е.Н. Крючкова
Согласовал	Зав. кафедрой «ПМ»	Е.Г. Боровцов
	руководитель направленности (профиля) программы	Е.Н. Крючкова

г. Барнаул

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

Код компетенции из УП и этап её формирования	Содержание компетенции	В результате изучения дисциплины обучающиеся должны:		
		знать	уметь	владеть
ОПК-1	Владением методологией теоретических и экспериментальных исследований в области профессиональной деятельности	Знать методы теоретических и экспериментальных исследований в области обработки естественных языков	Уметь применять методы теоретических и экспериментальных исследований в области построения автоматизированных систем обработки естественных языков	Владеть методами теоретических и экспериментальных исследований в области разработки автоматизированных систем обработки естественных языков
ОПК-2	Владением культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий	Знать принципы научных исследований в области обработки естественных языков	Уметь применять в процессе научных исследований в области построения автоматизированных систем обработки естественных языков основные принципы и подходы к моделированию синтаксиса и семантики языка	Владеть культурой анализа естественных языков
ПК-2	способность проводить комплексные исследования научных и технических проблем с применением современных технологий математического моделирования и вычислительного эксперимента	Знать научно-технические проблемы, современные технологии и математические модели, применяемые при анализе естественных языков	Уметь проводить вычислительные эксперименты при разработке моделей естественных языков	Владеть технологией научных исследований в области автоматической обработки естественных языков, в том числе технологией проведения вычислительных экспериментов при разработке и анализе моделей естественных языков

2. Место дисциплины в структуре образовательной программы

Дисциплины (практики), предшествующие изучению дисциплины, результаты освоения которых необходимы	Методы обработки результатов инженерного эксперимента в области математического моделирования, численных методов и комплексов программ, Обработка больших данных с помощью нейросетевых технологий

для освоения данной дисциплины.	
Дисциплины (практики), для которых результаты освоения данной дисциплины будут необходимы, как входные знания, умения и владения для их изучения.	Подготовка научно-квалификационной работы (диссертации) на соискание ученой степени кандидата наук

3. Объем дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося

Общий объем дисциплины в з.е. /час: 4 / 144
Форма промежуточной аттестации: Зачет

Форма обучения	Виды занятий, их трудоемкость (час.)				Объем контактной работы обучающегося с преподавателем (час)
	Лекции	Лабораторные работы	Практические занятия	Самостоятельная работа	
очная	0	0	18	126	18

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

Форма обучения: очная
Семестр: 6

Практические занятия (18ч.)

1. Введение в модели и алгоримы обработки естественных языков {беседа} (2ч.)[2,3] Природа сложности задач автоматической обработки текстов (АОТ) на естественном языке. Модели представления знаний.
Проблемы контекстной зависимости и многозначности.
Синтаксис и семантика естественного языка. Общие требования к семантической модели.
Статистические модели. Векторные представления.
Модели машинного обучения. Сверточные и рекуррентные нейросети. Архитектура LSTM.

2. Фаза предварительной обработки текста {метод кейсов} (2ч.)[2,9] Токенизацию. Удаление стоп-слов.
Приведение слов к единому регистру. Устранение шума.

Лемматизация. Обработка аббревиатур, сленга и коррекция ошибок.
3. Тезаурусы и лексиконы {метод кейсов} (2ч.)[4,6,7] Тезаурус WordNet. Классическая модель WordNet.
Семантические отношения. Синсеты. Русскоязычные тезаурусы РуТез и RuWordNet.
Лексикон. Тональный лексикон.
4. Статистические модели {метод кейсов} (4ч.)[2,3,4,5] Дистрибутивная гипотеза и понятие о близости значений слов.
Частотность. Взвешивание. Вес в пределах коллекции.
Векторные модели и плотные вектора в Word Embeddings.
Модели индексации: TF, TF-IDF, GloVe, Word2Vec.
Слово и его окрестность. Реализации word embeddings на основе метода GloVe, модели Skip-gram и CBOW (continuous Bag-of-words) в составе алгоритма Word2Vec.
Подходы к определению окрестности слова в модели:
предсказание слова по его окрестности в CBOW,
предсказание окрестности по слову в Skip-Gram.
5. Синтаксический и семантический анализ {метод кейсов} (4ч.)[1,2,4,8,9] Понятие грамматики. Контекстная зависимость. Контекстно-свободные грамматики.
Понятие дерева вывода. Естественные и формальные языки.
Семантический парсер естественного языка RML.
Ресурсы, применяемые в задачах АОТ.
Язык Питон и библиотеки обработки естемтаенных языков.
6. Модели в задаяах аннотирования текстов {беседа} (2ч.)[3,5] Общее и тематически-ориентированное аннотирование.
Извлекающие и генерирующие модели автоматического аннотирования.
Генерирующие методы на основе логико-семантических отношений между фрагментами текста.
Извлекающие методы на основе лексических цепочек, латентно-семантическом анализе, скрытом распределении Дирихле.
7. Модели в задачах классификации текстов {дискуссия} (2ч.)[2,6,7,9] Задача классификации. Понятие аспекта и категории.
Предобработка и индексация, уменьшение размерности пространства признаков, построение и обучение классификатора, оценка качества классификации.
Классификатор на основе метода логистической регрессии, наивного байесовского классификатора (NBC), k-ближайших соседей (KNN), метода опорных векторов (SVM), деревьев решений и случайных лесов.

Самостоятельная работа (126ч.)
. Самостоятельная работа(126ч.)[1,2,3,4,5,6,7,8,9] Самостоятельные научные исследования, в том числе с использованием

современных информационно-коммуникационных технологий

5. Перечень учебно-методического обеспечения самостоятельной работы обучающихся по дисциплине

Для каждого обучающегося обеспечен индивидуальный неограниченный доступ к электронно-библиотечным системам: Лань, Университетская библиотека он-лайн, электронной библиотеке АлтГТУ и к электронной информационно-образовательной среде:

1. Крючкова Е. Н. Основы теории конструирования компиляторов: Учебно-методическое пособие.- Барнаул: АлтГТУ, 2020. - 405с. Прямая ссылка: http://elib.altstu.ru/eum/download/pm/Kruchkova_OTKK_up.pdf Требует верификации

6. Перечень учебной литературы

6.1. Основная литература

2. Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 15.04.2021). — Режим доступа: для авториз. пользователей.

3. Юре, Л. Анализ больших наборов данных / Л. Юре, Р. Ананд, Д. У. Джеффри ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/93571 (дата обращения: 15.04.2021). — Режим доступа: для авториз. пользователей.

6.2. Дополнительная литература

4. Авдошин, С.М. Дискретная математика. Формально-логические системы и языки [Электронный ресурс] / С.М. Авдошин, А.А. Набебин. - Электрон. дан. - Москва : ДМК Пресс, 2018. - 390 с. - Режим доступа: https://e.lanbook.com/book/100912. - Загл. с экрана.

5. Мартин, О. Байесовский анализ на Python : руководство / О. Мартин ; перевод с английского А. В. Снастина. — Москва : ДМК Пресс, 2020. — 340 с. — ISBN 978-5-97060-768-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140585 (дата обращения: 15.04.2021). — Режим доступа: для авториз. пользователей.

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

6. http://wordnet.ru/

Русский WordNet

7. https://wordnet.princeton.edu/download/current-version

8. ANTLR [Электронный ресурс] // – Режим доступа: https://www.antlr.org/, свободный.

9. https://pythonist.ru/8-luchshih-bibliotek-obrabotki-estestvennogo-yazyka-dlya-python-nlp/

8. Фонд оценочных материалов для проведения текущего контроля успеваемости и промежуточной аттестации

Содержание промежуточной аттестации раскрывается в комплекте контролирующих материалов, предназначенных для проверки соответствия уровня подготовки по дисциплине требованиям ФГОС, которые хранятся на кафедре-разработчике РПД в печатном виде и в ЭИОС.

Фонд оценочных материалов (ФОМ) по дисциплине представлен в приложении А.

9. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем

Для успешного освоения дисциплины используются ресурсы электронной информационно-образовательной среды, образовательные интернет-порталы, глобальная компьютерная сеть Интернет. В процессе изучения дисциплины происходит интерактивное взаимодействие обучающегося с преподавателем через личный кабинет студента.

№пп	Используемое программное обеспечение
1	Acrobat Reader
2	Eclipse IDE
3	Java Runtime Environment
4	MASM32
5	Microsoft Office
6	Python
7	Qt Creator Open Source
8	Visual Studio
9	Windows
10	Антивирус Kaspersky

№пп	Используемые профессиональные базы данных и информационные справочные системы
1	IEEE Xplore - Интернет библиотека с доступом к реферативным и полнотекстовым статьям и материалам конференций. Бессрочно без подписки (https://ieeexplore.ieee.org/Xplore/home.jsp)
2	Springer - Издательство с доступом к реферативным и полнотекстовым материалам журналов и книг (https://www.springer.com/gp https://link.springer.com/)

10. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Наименование специальных помещений и помещений для самостоятельной работы
учебные аудитории для проведения занятий лекционного типа
учебные аудитории для проведения занятий семинарского типа
учебные аудитории для проведения курсового проектирования (выполнения курсовых работ)
учебные аудитории для проведения групповых и индивидуальных консультаций
учебные аудитории для проведения текущего контроля и промежуточной аттестации
помещения для самостоятельной работы
лаборатории
виртуальный аналог специально оборудованных помещений

Материально-техническое обеспечение и организация образовательного процесса по дисциплине для инвалидов и лиц с ограниченными возможностями здоровья осуществляется в соответствии с «Положением об обучении инвалидов и лиц с ограниченными возможностями здоровья».