Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет» (Новосибирский государственный университет, НГУ)

Гуманитарный институт

СОГЛАСОВАНО

Директор ГИ

_____ Зуев А.С.

«29» сентября 2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

МЕТОДЫ И АЛГОРИТМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Направление подготовки: 45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль): Математическая и прикладная лингвистика

Форма обучения: очная

Разработчики:

Старший преподаватель Бондаренко Иван Юрьевич _____

И.о. заведующего
кафедрой фундаментальной и прикладной лингвистики
д-р филос. наук, доцент Савостьянов А. Н. _____

Новосибирск

Содержание

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

Результаты освоения образовательной программы (компетенции)	Индикаторы	Результаты обучения по дисциплине
ПКС-1: Способен проводить научно-исследовательские и опытно-конструкторские разработки по отдельным разделам темы	ПКС-1.2. Демонстрирует знание математических понятий, теорий, методов в объеме достаточном для построения математических и компьютерных моделей естественных и формальных языков.	- знать основные методы машинного обучения, основанные на нейронных сетях и деревьях решений; - уметь решать задачи классификации текстов (в частности, анализа тональности), морфологического анализа, распознавания именованных сущностей с использованием методов машинного обучения; - владеть техникой разработки компьютерных моделей машинного обучения на языке программирования Python.

2. Место дисциплины в структуре образовательной программы

Дисциплина Методы и алгоритмы компьютерной лингвистики реализуется в 5 семестре в рамках Блока 1.

3. Трудоемкость 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

Трудоемкость дисциплины – 72 ч.
Форма промежуточной аттестации: 5 семестр – экзамен

№	Вид деятельности	Семестр
		5
1	Лекции, ч	16
2	Практические занятия, ч	16
3	Занятия в контактной форме, ч, из них	36
5	аудиторных занятий, ч	32
6	в электронной форме, ч	- -
7	консультаций, час.	2

8	промежуточная аттестация, ч	2
9	Самостоятельная работа, час.	36
10	Всего, ч	72

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

Лекции (16 ч)

Наименование темы и их содержание	Объем, час
1. Общая постановка задачи машинного обучения. Машинное обучение и искусственный интеллект. Виды машинного обучения: обучение с учителем, обучение без учителя, обучение с подкреплением	2
2. Задача классификации. Признаковое описание объекта. Дискриминантный и генеративный подходы к решению задачи классификации. Логистическая регрессия. Оценка качества алгоритма классификации. Проблема переобучения (overfitting) и методы борьбы с ней	2
3. Искусственные нейронные сети (ИНС). Модель нейрона, функция активации. Виды межнейронных связей. Проблема обучаемости и представляемости ИНС. Однослойный и многослойный персептроны.	2
4. Алгоритм обратного распространения ошибки для обучения многослойных персептронов и его модификации. Проблема регуляризации при обучении многослойного персептрона.	2
5. Векторные представления слов на основе двухслойного персептрона, обучающегося моделировать контекст фиксированной ширины. Модели CBOW и SkipGram. Модель FastText как развитие идеи векторных представлений слов для языков с богатой морфологией	2
6. Свёрточные нейронные сети для распознавания локально упорядоченных объектов. Архитектура и проблемы обучения. Алгоритмы классификации текстов на естественных языках с помощью свёрточных нейронных сетей	2
7. Рекуррентные нейронные сети для моделирования последовательностей. «Затухание градиента» и другие проблемы обучения рекуррентных нейронных сетей. Преимущества рекуррентных нейронных сетей для обработки текстов произвольного размера. Рекуррентные сети с долгосрочной и краткосрочной памятью (Long Short Term Memory, LSTM). Применение таких сетей для морофологического анализа, распознавания именованных сущностей в текстах и языкового моделирования. Алгоритм ELMo для генерации векторных представлений слов с учётом «глубокого» контекста.	2
8. Алгоритмы деревьев решений и методы коллективного распознавания как ещё один подход к задаче классификации текстов, альтернативный нейронным сетям. Алгоритмы C4.5, случайного леса и градиентного бустинга	2

Практические занятия (16 ч)

Содержание практического занятия	Объем, час
Семинар по теме «Общая постановка задачи машинного обучения. Машинное обучение и искусственный интеллект. Виды машинного обучения: обучение с учителем, обучение без учителя, обучение с подкреплением»	2
Семинар по теме «Задача классификации. Признаковое описание объекта. Дискриминантный и генеративный подходы к решению задачи классификации. Логистическая регрессия. Оценка качества алгоритма классификации. Проблема переобучения (overfitting) и методы борьбы с ней»	2
Семинар по теме «Искусственные нейронные сети (ИНС). Модель нейрона, функция активации. Виды межнейронных связей. Проблема обучаемости и представляемости ИНС. Однослойный и многослойный персептроны»	2
Семинар по теме «Алгоритм обратного распространения ошибки для обучения многослойных персептронов и его модификации. Проблема регуляризации при обучении многослойного персептрона»	2
Семинар по теме «Векторные представления слов на основе двухслойного персептрона, обучающегося моделировать контекст фиксированной ширины. Модели CBOW и SkipGram. Модель FastText как развитие идеи векторных представлений слов для языков с богатой морфологией»	2
Семинар по теме «Свёрточные нейронные сети для распознавания локально упорядоченных объектов. Архитектура и проблемы обучения. Алгоритмы классификации текстов на естественных языках с помощью свёрточных нейронных сетей»	2
Семинар по теме «Рекуррентные нейронные сети для моделирования последовательностей. «Затухание градиента» и другие проблемы обучения рекуррентных нейронных сетей. Преимущества рекуррентных нейронных сетей для обработки текстов произвольного размера. Рекуррентные сети с долгосрочной и краткосрочной памятью (Long Short Term Memory, LSTM). Применение таких сетей для морофологического анализа, распознавания именованных сущностей в текстах и языкового моделирования. Алгоритм ELMo для генерации векторных представлений слов с учётом «глубокого» контекста»	2
Семинар по теме «Алгоритмы деревьев решений и методы коллективного распознавания как ещё один подход к задаче классификации текстов, альтернативный нейронным сетям. Алгоритмы C4.5, случайного леса и градиентного бустинга»	2

Самостоятельная работа студентов (36 ч)

Перечень занятий на СРС	**Объем, час**
Самостоятельное освоение лекционного материала	10
Подготовка к семинарским занятиям	11

Подготовка к экзамену	15
ИТОГО	36

5. Перечень учебной литературы

5.1. Основная литература

1. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб.пособие— М.: МИЭМ, 2011. — 272 с. URL: http://e-lib.nsu.ru/dsweb/Get/Resource-1583/page001.pdf.
2. Батура Т.В. Математическая лингвистика и автоматическая обработка текстов на естественном языке : учебное пособие : [для студентов и аспирантов ФИТ, ММФ и ГФ (отделение фундаментальной и прикладной лингвистики) НГУ] / Т.В. Батура ; М-во образования и науки Рос. Федерации, Новосиб. гос. ун-т, Фак. информ. технологий, Каф. систем информатики .— Новосибирск : Редакционно-издательский центр НГУ, 2016 .— 165 с. : ил. ; 20 см. .— Библиогр. в конце глав .— URL: http://e-lib.nsu.ru/dsweb/Get/Resource-1583/page001.pdf.
3. Гладкий, Алексей Всеволодович. Элементы математической лингвистики / А. В. Гладкий, И. А. Мельчук. Москва : Наука, 1969. 192 с. (20 экземпляров).

6. Перечень учебно-методических материалов по самостоятельной работе обучающихся

4. Касьянов В.Н. Лекции по теории формальных языков, автоматов и сложности вычислений : [учебное пособие для студентов-математиков и информатиков] / В.Н. Касьянов ; Гос. ком. Рос. Федерации по высш. образованию, Новосиб. гос. ун-т. Новосибирск : Редакционно-издательский отдел НГУ, 1995. 112 с. : ил. (186 экземпляров).

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

7.1.Ресурсы сети «Интернет»

Освоение дисциплины используются следующие ресурсы:

- электронная информационно-образовательная среда НГУ (ЭИОС);
- образовательные интернет-порталы;
- информационно-телекоммуникационная сеть Интернет.

Взаимодействие обучающегося с преподавателем (синхронное и (или) асинхронное) осуществляется через личный кабинет студента в ЭИОС, электронную почту.

7.2 Современные профессиональные базы данных:

Не используется

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

8.1 Перечень программного обеспечения

Для обеспечения реализации дисциплины используется стандартный комплект программного обеспечения (ПО), включающий регулярно обновляемое лицензионное ПО Windows и MS Office.

8.2 Информационные справочные системы

№	*Наименование*	*Назначение*
1	Python	Язык программирования
2	NLTK	Программная библиотека для компьютерной лингвистики
3	Tensorflow и Keras	Программные библиотеки для моделирования нейронных сетей
4	Scikit-Learn	Программная библиотека для моделирования «классических» методов машинного обучения (деревьев решений, опорных векторов и т.п.)
5	Gensim	Программная библиотека для моделирования дистрибутивной семантики слов и текстов

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Для реализации дисциплины Методы и алгоритмы компьютерной лингвистики используются специальные помещения:

1. Учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля, промежуточной и итоговой аттестации;

2. Помещения для самостоятельной работы обучающихся.

Учебные аудитории укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду НГУ.

Для проведения занятий лекционного типа предлагаются следующие наборы демонстрационного оборудования и учебно-наглядных пособий:

- комплект лекций-презентаций по темам дисциплины.

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья

осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

Реализация дисциплины может осуществляться с применением дистанционных образовательных технологий.

10. Оценочные средства для проведения текущего контроля и промежуточной аттестации по дисциплине

10.1 Порядок проведения текущего контроля и промежуточной аттестации по дисциплине

Текущий контроль успеваемости:

Текущий контроль проводится по посещаемости, участию в работе на аудиторных занятиях.

Промежуточная аттестация:

Промежуточная аттестация по дисциплине проводится в форме зачёта. Результаты прохождения аттестации оцениваются по шкале «неудовлетворительно», «удовлетворительно», «хорошо», «отлично». Оценки «отлично», «хорошо», «удовлетворительно», означают успешное прохождение промежуточной аттестации.

Минимальная положительная оценка «удовлетворительно» ставится студенту, если он владеет теоретическим материалом, допуская существенные ошибки по содержанию рассматриваемых (обсуждаемых) вопросов, испытывает затруднения в формулировке собственных суждений, допускает значительные ошибки при ответе на дополнительные вопросы.

На подготовку к ответу отводится 30 минут. Преподаватель может задавать дополнительные вопросы по всем темам курса (случайная выборка). Оценка сообщается в тот же день.

Описание критериев и шкал оценивания индикаторов достижения результатов обучения по дисциплине

Таблица 10.1

Код компетенции	Индикатор	Результат обучения по дисциплине	Оценочное средство
ПКС-1	ПКС-1.2. Демонстрирует знание математических понятий, теорий, методов в объеме достаточном для построения математических и компьютерных моделей естественных и	знать основные методы машинного обучения, основанные на нейронных сетях и деревьях решений; - уметь решать задачи классификации текстов (в частности, анализа	Вопросы к экзамену

	формальных языков.	тональности), морфологического анализа, распознавания именованных сущностей с использованием методов машинного обучения; - владеть техникой разработки компьютерных моделей машинного обучения на языке программирования Python.	

Таблица 10.2

Критерии оценивания результатов обучения	**Шкала оценивания**
<u>Экзамен:</u> – Студент владеет теоретическим и практическим материалом, - формулирует собственные, самостоятельные, обоснованные, аргументированные суждения, - представляет полные и развернутые ответы на дополнительные вопросыПри изложении ответа на вопрос(ы) экзаменационного билета обучающийся мог допустить непринципиальные неточности.	*Отлично*
<u>Экзамен:</u> - Студент в основном владеет теоретическим материалом, - формулирует собственные, самостоятельные, обоснованные, аргументированные суждения, - допускает незначительные ошибки при ответе на дополнительные вопросы.	*Хорошо*
<u>Экзамен:</u> - владеет теоретическим материалом, допуская существенные ошибки по содержанию рассматриваемых вопросов, - испытывает затруднения в формулировке собственных суждений, - допускает значительные ошибки при ответе на дополнительные вопросы	*Удовлетворительно*
<u>Экзамен:</u> – фрагментарное и недостаточное владение теоретическим материалом, – неспособность сформулировать собственные рассуждения, – отсутствие ответов на дополнительные вопросы.	*Неудовлетво-рительно*

Типовые контрольные задания и иные материалы, необходимые для оценки результатов обучения

1. Разработать алгоритм классификации тональности твита, используя материалы соревнования SentiRuEval-2016. Необходимо использовать свёрточную нейронную сеть на Keras или PyTorch и предобученные word2vec из RDT https://nlpub.ru/RDT или RusVectores http://rusvectores.org/ru/ .
2. Разработать систему распознавания именованных сущностей трёх классов: имена людей (PER, или PERSON), названия организаций (ORG, или ORGANIZATION) и геолокаций (LOC, или LOCATION). Задание соответствует первой дорожке соревнования FactRuEval-2016.
3. Разработать алгоритм классификации тональности твита, используя глубокую свёрточную нейронную сеть на Keras или на PyTorch и посимвольное представление текста.

Лист актуализации рабочей программы дисциплины
«Математические модели языка»

№	Характеристика внесенных изменений (с указанием пунктов документа)	Дата и № протокола ученого совета Гуманитарного института	Подпись ответственного