МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
высшего образования - программы магистратуры
по направлению подготовки
09.04.01 Информатика и вычислительная техника,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Обработка естественного языка

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Искусственный интеллект и предиктивная аналитика в транспортных системах

Форма обучения: Очная

Рабочая программа дисциплины (модуля) в виде электронного документа выгружена из единой корпоративной информационной системы управления университетом и соответствует оригиналу

Простая электронная подпись, выданная РУТ (МИИТ)
ID подписи: 5665
Подписал: заведующий кафедрой Нутович Вероника Евгеньевна
Дата: 01.09.2024

1. Общие сведения о дисциплине (модуле).

Цель дисциплины «Обработка естественного языка» в освоении принципов интеллектуального анализа текста в системах машинного обучения.

В рамках дисциплины формируются знания о компьютерной лингвистике, способах предварительной обработки и преобразования языка, определении связи между словами, построении моделей классификации текста и архитектурах нейронных сетей для обработки текстов на естественном языке.

На лабораторных занятиях у обучающихся формируются навыки работы с современными библиотеками интеллектуального анализа и обработки естественного языка PyTorch, TensorFlow языка программирования Python.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

ПК-3 - Способен спроектировать, разработать, обучить, оценить и развернуть модели искусственного интеллекта в соответствии с методологией MLOps.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

Уметь:

- проводить автомазитизрованную обработку ествественного языка;
- представлять текстовую информацию в векторной плоскости представления слов и связанных выражений;
- оценивать и нормализировать размерность векторного отображения слов и связанных выражений;

Знать:

- принципы текстовой классификации;
- типы, задачи и особенности построения различных нейронных сетей;
- отличительные особенности задач классификации для машинного перевода и извлечения ключевых слов.

Владеть:

- навыком предварительной обработки и построения семантических связей между словами;
- навыком построения рекуррентных нейронных сетей для анализа текста

на естественном языке;

- навыком интегрирования и развертывания среды разработки и обучения модели искусственного интеллекта для анализа текста на естественном языке.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 4 з.е. (144 академических часа(ов).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Тип учебных занятий	Количество часов	
	Всего	Сем. №3
Контактная работа при проведении учебных занятий (всего):	36	36
В том числе:		
Занятия лекционного типа	18	18
Занятия семинарского типа	18	18

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 108 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Обработка естественного языка. Рассматриваемые вопросы: - понятие NLP (natural language processing) и его основные задачи: классификация, машинный перевод, извлечение ключевых слов (NER), информационный поиск, извлечение информации; - сложности компьютерной обработки естественного языка.
2	Естественный язык как данные. Рассматриваемые вопросы: - признаки языка: лингвистические, контекстные и структурные; - предварительная обработка текста: сегментация, лексемизация, маркировка частей речи, анализ настроений; - преобразование текста: словоизменение и лемматизация, орфографическая коррекция.
3	Текстовая классификация. Рассматриваемые вопросы: - инвертированный индекс; - фрагментное индексирование: n-граммы.
4	Векторизация текста. Рассматриваемые вопросы: - построение векторного набора слов: быстрое кодирование (One-Hot Encoding), вещественные векторы; - методы «мешка слов»: Bag of Words(BoW), Continuous Bag of Words (CBow), Deep CBow; - модель скип-грамм (Skip Gram); - статистические меры: частота слова (TF), TF/IDF; - семантические вложения (эмбеддинги): Word2Vec и GloVe библиотеки;
5	Рекуррентные нейросети. Рассматриваемые вопросы: - рекурентные нейросети (RNN) и особенности ее функционирования; - двунаправленные и многослойные RNN.
6	Генеративные рекуррентные нейросети. Рассматриваемые вопросы: - генерация и создание текста нейронной сетью; - трансформерные модели.
7	Машинный перевод как одна из задач обработки естественного языка. Рассматриваемые вопросы: - машинный перевод на базе лингвистических правил; - статистический машинный перевод; - нейронный машинный перевод.
8	Извлечение ключевых слов как одна из задач обработки естественного языка. Рассматриваемые вопросы: - построение токенов и словари именнованных токенов (gazetteers); - модели классификации токенов: онтологическая инженерия и глубокое обучение.
9	Автоматическая обработка языка. Рассматриваемые вопросы: - преобразование аудио в текст и обратное воспроизведение; - предварительная обработка текста; - грамматические модели русского языка в контекста автоматической обработки. - когнитивные сервисы для работы с языком: Google Natural Language, Azure Cognitive Services, TextRazor, AssemblyAI, Dandelion, Allganize.

4.2. Занятия семинарского типа.

Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	Классификация текста. В результате выполнения лабораторных работ студент получает навыки разделения текстовых данных на заданные классы категорий с помощью библитек PyTorch и TensorFlow языка программирования Python.
2	Сегментация текста. В результате выполнения лабораторных работ студент получает навыки выделения связанных слов (токенов) на наботе текстовых данных с помощью библитек PyTorch и TensorFlow языка программирования Python.
3	Векторизация текста. В результате выполнения лабораторных работ студент получает навыки построения словаря токенов из набора текстовых данных с помощью библитек PyTorch и TensorFlow языка программирования Python.
4	Фрагментное индексирование текста. В результате выполнения лабораторных работ студент получает навыки выделения связанных выражений, состоящих из двух, трех и n-слов, фрагментированного набора текстовых данных с помощью библитек PyTorch и TensorFlow языка программирования Python.
5	Расчет векторов BoW текста. В результате выполнения лабораторных работ студент получает навыки уменьшения размерности векторного отображения словаря через построение векторных представлений Bow, CBow и DeepCBow с помощью библитек PyTorch и TensorFlow языка программирования Python.
6	Расчет частоты слов и фраз текста. В результате выполнения лабораторных работ студент получает навыки взвешенного уменьшения размерности векторного отображения словаря с учетом частотно веса для разных слов с помощью библитек PyTorch и Scikit Learn, Keras языка программирования Python.
7	Семантические вложения. В результате выполнения лабораторных работ студент получает навыки использования встраивания слов для уменьшения размерности векторов позиционного представления с помощью библитек Word2Vec или GloVe языка программирования Python.
8	Использование рекуррентных нейронных сетей. В результате выполнения лабораторных работ студент получает навыки использования рекуррентных нейронных сетей для анализа текста с помощью библитек PyTorch и TensorFlow языка программирования Python.
9	Использование рекуррентных нейронных сетей. В результате выполнения лабораторных работ студент получает навыки использования рекуррентных нейронных сетей для создания текста с помощью библитек PyTorch и TensorFlow языка программирования Python.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.
2	Подготовка к лабораторным работам.

3	Подготовка к промежуточной аттестации.
4	Подготовка к текущему контролю.

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Бенджамин Бенгфорт. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – Санкт-Петербург: Питер, 2021. – 368 с. – ISBN 978-5-4461-1153-4.	https://ibooks.ru/bookshelf/365298/reading (дата обращения: 1.11.2022). – Текст : электронный
2	Лейн Хобсон. Обработка естественного языка в действии. - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-1371-2.	https://ibooks.ru/bookshelf/371695/reading (дата обращения: 1.11.2022). – Текст : электронный
3	Брайан Макмахан. Знакомство с PyTorch: глубокое обучение при обработке естественного языка. - Санкт-Петербург : Питер, 2021. - 256 с. - ISBN 978-5-4461-1241-8.	https://ibooks.ru/bookshelf/374453/reading (дата обращения: 1.11.2022). – Текст : электронный
4	Ганегедара, Т. Обработка естественного языка с TensorFlow : руководство / Т. Ганегедара ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5.	https://e.lanbook.com/book/140584 (дата обращения: 1.11.2022). – Текст : электронный
5	Гольдберг Й. Нейросетевые методы в обработке естественного языка / пер. с анг. А. А. Слинкина / Й. Гольдберг. - Москва : ДМК Пресс, 2019. - 282 с. - ISBN 978-5-97060-754-1.	https://ibooks.ru/bookshelf/385120/reading (дата обращения: 1.11.2022). – Текст : электронный
6	Бринк Х. Машинное обучение / Х. Бринк, Д. Ричардс, М. Феверолф. – Санкт-Петербург : Питер, 2017. – 336 с. – ISBN 978-5-496-02989-6.	https://ibooks.ru/bookshelf/355472/reading (дата обращения: 1.11.2022). – Текст : электронный

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

Электронно-библиотечная система Научно-технической библиотеки РУТ(МИИТ) (http://library.miit.ru/)

Электронно-библиотечная система издательства «Лань»

(http://e.lanbook.com/)

Электронно-библиотечная система ibooks.ru (http://ibooks.ru/)

Электронная документация по когнитивным сервисам в MS Azure (https://learn.microsoft.com/ru-ru/azure/cognitive-services/cognitive-services-and-machine-learning)

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Прикладное программное обеспечение

Браузер Microsoft Internet Explorer или его аналоги

Пакет офисных программ Microsoft Office или его аналоги

Среда разработки PyCharm Community Edition

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для практических занятий – наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Экзамен в 3 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

заведующий кафедрой, доцент, к.н. кафедры «Цифровые технологии управления транспортными процессами» В.Е. Нутович

старший преподаватель кафедры «Цифровые технологии управления транспортными процессами» Е.А. Заманов

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической комиссии Н.А. Андриянова