

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»



УТВЕРЖДАЮ

Директор по образовательной деятельности

С.Т. Князев

С.Т. Князев

«12»

октября

2021 г.

РАБОЧАЯ ПРОГРАММА МОДУЛЯ

Код модуля
М.1.9

Модуль
Анализ естественного языка

Екатеринбург, 2021

Перечень сведений о рабочей программе модуля	Учетные данные
Образовательная программа Прикладной искусственный интеллект (Practical Artificial Intelligence)	Код ОП 09.04.02
Направление подготовки Информационные системы и технологии	Код направления и уровня подготовки 09.04.02

Области образования, в рамках которых реализуется модуль образовательной программы по СУОС УрФУ:

№ п/п	Перечень областей образования, для которых разработан СУОС УрФУ	Уровень подготовки
1.	Инженерное дело, технологии и технические науки	магистратура

Программа модуля составлена авторами:

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	кандидат технических наук, нет	доцент	Кафедра информационных технологий и систем управления, ИРИТ-РТФ, УрФУ
2	Аксёнов Александр Сергеевич	нет	инженер	Отдел обеспечения деятельности ИРИТ-РТФ

Рекомендовано учебно-методическим советом института радиоэлектроники и информационных технологий - РтФ

Протокол № 7 от 11 октября 2021 г.

1. ОБЩАЯ ХАРАКТЕРИСТИКА МОДУЛЯ АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА

1.1. Аннотация содержания модуля

Модуль состоит из одной дисциплины «Анализ естественного языка».

Дисциплина «Анализ естественного языка» знакомит студентов с современными методами анализа естественного языка, основанными на глубоких нейронных сетях и машинном обучении. Рассматриваются задачи классификации текста, автоматической генерации текста с использованием рекуррентных нейронных сетей, включая LSTM и GRU, одномерных сверточных сетей, а также сетей с архитектурой Transformer.

1.2. Структура и объем модуля

Таблица 1

№ п/п	Перечень дисциплин модуля в последовательности их освоения	Объем дисциплин модуля и всего модуля в зачетных единицах и часах
1.	Анализ естественного языка	6 з.е./216 ч.
ИТОГО по модулю:		6 з.е./216 ч.

1.3. Последовательность освоения модуля в образовательной программе

Пререквизиты модуля	нет
Постреквизиты и корреквизиты модуля	нет

1.4. Распределение компетенций по дисциплинам модуля, планируемые результаты обучения (индикаторы) по модулю

Таблица 2

Перечень дисциплин модуля	Код и наименование компетенции	Индикаторы достижения компетенции	Планируемые результаты обучения
1	2	3	4
Анализ естественного языка	ОПК-8. Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта	ОПК-8.2. Разрабатывает оригинальные программные средства для решения задач в области создания и применения искусственного интеллекта	ОПК-8.2. 3-1. Знает принципы разработки оригинальных программных средств для решения профессиональных задач ОПК-8.2. У-1. Умеет разрабатывать оригинальные программные средства для решения задач в области создания и применения искусственного интеллекта

	ПК-7. Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях	ПК-7.2. Руководит проектами в области сквозной цифровой субтехнологии «Обработка естественного языка»	ПК-7.2. З-1. Знает принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка» ПК-7.2. У-1. Умеет руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»
--	---	---	--

1.5. Форма обучения

Обучение по дисциплинам модуля может осуществляться в очной форме

2. СОДЕРЖАНИЕ И ОБЕСПЕЧЕНИЕ РЕАЛИЗАЦИИ ДИСЦИПЛИН МОДУЛЯ

**ПРОГРАММА МОДУЛЯ
АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА**

**РАЗДЕЛ 2. СОДЕРЖАНИЕ И ОБЕСПЕЧЕНИЕ РЕАЛИЗАЦИИ ДИСЦИПЛИН
МОДУЛЯ**

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ 1
АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА**

Рабочая программа дисциплины составлена авторами:

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	кандидат технических наук, нет	доцент	Кафедра информационных технологий и систем управления, ИРИТ-РТФ, УрФУ
2	Аксёнов Александр Сергеевич	нет	инженер	Отдел обеспечения деятельности ИРИТ-РТФ

Рекомендовано учебно-методическим советом института радиоэлектроники и информационных технологий - РТФ

Протокол № 7 от 11 октября 2021 г.

2. СОДЕРЖАНИЕ И ОСОБЕННОСТИ РЕАЛИЗАЦИИ ДИСЦИПЛИНЫ 1 АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА

2.1. Технологии обучения, используемые при изучении дисциплины модуля

- Традиционная (репродуктивная) технология;
- Смешанное обучение с использованием онлайн-курса;
- Исключительно электронное обучение с использованием онлайн-курса.

2.2. Содержание дисциплины

Таблица 1

Код раздела, темы	Раздел, тема дисциплины*	Содержание
1.	Теоретические аспекты обработки естественного языка.	Синтаксический, морфологический, семантический и графематический анализ, омонимия, задачи лингвистического анализа
2.	Предварительная обработка текста.	Очистка текста, токенизация, стемминг, лемматизация, удаление стоп-слов, фильтрация наиболее частотных и наименее частотных слов.
3.	Векторизация текста.	Построение словаря, мешок слов, TF-IDF, word2vec, fasttext, LDA, LSI, GloVe.
4.	Машинное обучение для обработки текстов.	Решение задач классификации и определения тональности методами классического машинного обучения на основе векторных моделей.
5.	Нейронные сети в решении задач текстовой обработки.	Архитектуры нейронных сетей для обработки текстов: рекуррентные (LSTM, GRU), одномерные сверточные. Применение нейронных сетей для обработки текстов.
6.	Языковая модель.	Языковая модель и дистрибутивная семантика. Обучение векторной модели. Задача генерации текста. Различные подходы к генерации текста.
7.	Поиск именованных существностей.	Задача поиска именованных существностей в тексте. Применение нейронных сетей для поиска именованных существностей.
8.	Механизм внимания. Трансформер.	Механизм внимания в нейронных сетях. Применение механизма внимания для обработки текста. Нейронные сети с архитектурой Transformer. Нейронные сети BERT, GPT. Перенос обучения.

2.3. Программа дисциплины реализуется полностью на иностранном языке.

2.4. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ Анализ естественного языка

Электронные ресурсы (издания)

1. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. URL: <http://www.machinelearning.ru/> (дата обращения: 04.10.2021).
2. Цитульский Антон Максимович, Иванников Александр Владимирович, Рогов Илья Сергеевич NLP - Обработка естественных языков // StudNet. 2020. №6. URL: <https://cyberleninka.ru/article/n/nlp-obrabotka-estestvennyh-yazykov> (дата обращения:

04.10.2021).

3. Чернобаев Игорь Дмитриевич, Суркова Анна Сергеевна, Панкратова Анна Зурабовна Моделирование текстов с использованием рекуррентных нейронных сетей // Труды НГТУ им. П. Е. Алексеева. 2018. №1 (120). URL: <https://cyberleninka.ru/article/n/modelirovanie-tekstov-s-ispolzovaniem-rekurrentnyh-neyronnyh-setey> (дата обращения: 04.10.2021).

4. Браславский П.И. Введение в обработку естественного языка. URL: <https://stepik.org/course/1233/> (дата обращения: 04.10.2021).

5. Роман Суворов, Анастасия Янина, Алексей Сильвестров, Николай Капырин. Нейронные сети и обработка текста URL: <https://stepik.org/course/54098> (дата обращения: 04.10.2021).

6. Онлайн курс “Программирование глубоких нейронных сетей на Python”. URL: <https://openedu.ru/course/urfu/PYDNN/> (дата обращения: 05.10.2021).

7. Онлайн курс “Generating discrete sequences: language and music”. URL: <https://www.edx.org/course/generating-discrete-sequences-language-and-music> (дата обращения: 05.10.2021).

Профессиональные базы данных, информационно-справочные системы

1. Цифровая библиотека научно-технических изданий Института инженеров по электротехнике и радиоэлектронике (Institute of Electrical and Electronic Engineers (IEEE)) на английском языке – <http://www.ieee.org/ieeexplore>

2. Oxford University Press – <http://www.oxfordjournals.org/en/>

3. Архив препринтов с открытым доступом – <https://arxiv.org/>

Материалы для лиц с ОВЗ

Весь контент ЭБС представлен в виде файлов специального формата для воспроизведения синтезатором речи, а также в тестовом виде, пригодном для прочтения с использованием экранной лупы и настройкой контрастности.

Базы данных, информационно-справочные и поисковые системы

1. Academic Search Ultimate EBSCO publishing – <http://search.ebscohost.com>

2. eBook Collections Springer Nature – <https://link.springer.com/>

3. Гугл Академия – <https://scholar.google.ru/>

4. Электронный научный архив УрФУ <https://elar.urfu.ru/>

5. Зональная научная библиотека (УрФУ) - <http://lib.urfu.ru/>

6. Портал информационно-образовательных ресурсов УрФУ <https://study.urfu.ru/>

7. Электронно-библиотечная система «Лань» – <https://e.lanbook.com/>

8. Университетская библиотека ONLINE – <https://biblioclub.ru/>

9. Электронно-библиотечная система "Библиокомплектатор" (IPRbooks)

<http://www.bibliocomplectator.ru/available>

10. Электронные информационные ресурсы Российской государственной библиотеки <https://www.rsl.ru/>

11. Научная электронная библиотека «КиберЛенинка» <https://cyberleninka.ru/>

2.5. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ Анализ естественного языка

Сведения об оснащённости дисциплины специализированным и лабораторным оборудованием и программным обеспечением

Таблица 2

№ п/п	Виды занятий	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа
1.	Лекции; Лабораторные занятия	Компьютерный класс. Мультимедийный проектор с экраном; Сетевое оборудование; Локальная сеть с выходом в глобальную сеть Интернет.	Используется бесплатно-распространяемое программное обеспечение: 1. Python – https://www.python.org/ 2. TensorFlow – https://www.tensorflow.org/ 3. Hugging Face – https://huggingface.co/ 4. Веб - среда разработки для языка программирования Python: Google Colab – https://colab.research.google.com/

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

Код модуля
М.1.9

Модуль
Анализ естественного языка

Екатеринбург, 2021

Оценочные материалы по модулю составлены авторами:

№ п/п	Фамилия, имя, отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	кандидат технически х наук, нет	доцент	Кафедра информационных технологий и систем управления, ИРИТ- РТФ, УрФУ

1. СТРУКТУРА И ОБЪЕМ МОДУЛЯ Анализ естественного языка

№ п/п	Перечень дисциплин модуля в последовательности их освоения	Объем дисциплин модуля и всего модуля в зачетных единицах и часах	Форма итоговой промежуточной аттестации по дисциплинам модуля и в целом по модулю
1.	Анализ естественного языка	6 / 216	<i>Зачет (3 семестр)</i>
ИТОГО по модулю:		6 / 216	

2. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО МОДУЛЮ

Не предусмотрено.

Раздел 3. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ 1
АНАЛИЗ ЕСТЕСТВЕННОГО ЯЗЫКА

Модуль М.1.9 Анализ естественного языка

Оценочные материалы составлены автором:

№ п/п	Фамилия Имя Отчество	Ученая степень, ученое звание	Должность	Подразделение
1	Созыкин Андрей Владимирович	кандидат технических наук, нет	доцент	Кафедра информационных технологий и систем управления, ИРИТ-РТФ, УрФУ

1. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ (ИНДИКАТОРЫ) ПО ДИСЦИПЛИНЕ МОДУЛЯ Анализ естественного языка

Индикатор – это признак / сигнал/ маркер, который показывает, на каком уровне обучающийся должен освоить результаты обучения и их предъявление должно подтвердить факт освоения предметного содержания данной дисциплины, указанного в табл. 1.3 РПМ-РПД.

Индикаторы должны учитываться при выборе и составлении заданий контрольно-оценочных мероприятий (оценочных средств) текущей и промежуточной аттестации.

Таблица 1

Код и наименование компетенций, формируемые с участием дисциплины	Индикаторы достижения компетенции	Планируемые результаты обучения	Контрольно-оценочные средства для оценивания достижения результата обучения по дисциплине
1	2	3	4
ОПК-8. Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта	ОПК-8.2. Разрабатывает оригинальные программные средства для решения задач в области создания и применения искусственного интеллекта	ОПК-8.2. З-1. Знает принципы разработки оригинальных программных средств для решения профессиональных задач ОПК-8.2. У-1. Умеет разрабатывать оригинальные программные средства для решения задач в области создания и применения искусственного интеллекта	Контрольные работы (2 шт.) Домашние работы (2 шт.) Выполнение практических работ Зачет
ПК-7. Способен руководить проектами по созданию, внедрению и использованию одной или нескольких сквозных цифровых субтехнологий искусственного интеллекта в прикладных областях	ПК-7.2. Руководит проектами в области сквозной цифровой субтехнологии «Обработка естественного языка»	ПК-7.2. З-1. Знает принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка» ПК-7.2. У-1. Умеет руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта на основе сквозной цифровой субтехнологии «Обработка естественного языка»	Контрольные работы (2 шт.) Домашние работы (2 шт.) Выполнение практических работ Зачет

2. ВИДЫ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ, ВКЛЮЧАЯ МЕРОПРИЯТИЯ ТЕКУЩЕЙ АТТЕСТАЦИИ

2.1. Распределение объема времени по видам учебной работы

Таблица 2

№ п/ п	Наименование дисциплины модуля Анализ естественного языка	Объем времени, отведенный на освоение дисциплины модуля								
		Аудиторные занятия, час.				Промежуточная аттестация (форма итогового контроля)	Контактная работа (час.)	Самостоятельная работа студента, включая текущую аттестацию (час.)	Всего по дисциплине	
		Занятия лекцион ного типа	Практиче ские работы	Лаборато рные работы	Всего				Час.	Зач. ед.
1	2	3	4	5	6	7	8	9	10	11
1.	Анализ естественного языка	36	0	72	108	Зачет	124,45	91,55	216	6

2.1. Виды СРС, количество и объем времени на контрольно-оценочные мероприятия СРС по дисциплине

Контрольно-оценочные мероприятия СРС включают самостоятельное изучение материала, подготовку к аудиторным занятиям и мероприятиям текущего контроля, выполнение и оформление внеаудиторных мероприятий текущего контроля и подготовку к мероприятиям промежуточного контроля.

Таблица 3

№ п/п	Вид самостоятельной работы студента по дисциплине модуля	Количество контрольно-оценочных мероприятий СРС	Объем контрольно-оценочных мероприятий СРС (час.)
1.	Подготовка к аудиторным занятиям и мероприятиям текущего контроля: лекционным, лабораторным занятиям.		27 час.
2.	Выполнение и оформление мероприятий текущего контроля:		
2.1	Контрольная работа	2	10 час.
2.2	Домашняя работа №1	1	5 час.
2.3	Домашняя работа №2	1	5 час.
3.	Подготовка к зачету	зачет (3 семестр)	12 час.
4.	Самостоятельное изучение материала		32,55 час.
Итого на СРС по дисциплине:			91,55 час.

3. ПРОЦЕДУРЫ КОНТРОЛЯ И ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ В РАМКАХ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ МОДУЛЯ В БАЛЬНО-РЕЙТИНГОВОЙ СИСТЕМЕ (ТЕХНОЛОГИЧЕСКАЯ КАРТА БРС)

Процедуры текущей и промежуточной аттестации по дисциплине

1.Лекции: коэффициент значимости совокупных результатов лекционных занятий – 0.5		
Текущая аттестация на лекциях	Сроки – семестр, учебная неделя	Максимальная оценка в баллах
<i>Контрольная работа</i>	2 сем.	70
<i>Самостоятельное изучение материала</i>	2 сем.	30
Весовой коэффициент значимости результатов текущей аттестации по лекциям – 0.5		
Промежуточная аттестация по лекциям – Зачет		
Весовой коэффициент значимости результатов промежуточной аттестации по лекциям – 0.5		
2. Практические/семинарские занятия: Не предусмотрены коэффициент значимости совокупных результатов практических занятий – 0		
3.Лабораторные занятия коэффициент значимости совокупных результатов практических/семинарских занятий – 0.5		
Текущая аттестация на практических/семинарских занятиях	Сроки – семестр, учебная неделя	Максимальная оценка в баллах
<i>Выполнение и оформление лабораторных работ</i>	2 сем.	50
<i>Домашняя работа №1</i>	2 сем.	25
<i>Домашняя работа №2</i>	2 сем.	25
Весовой коэффициент значимости результатов текущей аттестации по практическим/семинарским занятиям – 1		
Промежуточная аттестация по практическим/семинарским занятиям– не предусмотрена		
Весовой коэффициент значимости результатов промежуточной аттестации по практическим/семинарским занятиям – 0		

4. КРИТЕРИИ И УРОВНИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ МОДУЛЯ

4.1.В рамках БРС применяются утвержденные на кафедре/институте критерии (признаки) оценивания достижений студентов по дисциплине модуля (табл. 4) в рамках контрольно-оценочных мероприятий на соответствие указанным в табл.1 результатам обучения (индикаторам).

Таблица 4

Критерии оценивания учебных достижений обучающихся

Результаты обучения	Критерии оценивания учебных достижений, обучающихся на соответствие результатам обучения/индикаторам
Знания	Студент демонстрирует знания и понимание в области изучения на уровне указанных индикаторов и необходимые для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Умения	Студент может применять свои знания и понимание в контекстах, представленных в оценочных заданиях, демонстрирует освоение умений на уровне указанных индикаторов и необходимых для продолжения обучения и/или выполнения трудовых функций и действий, связанных с профессиональной деятельностью.
Опыт /владение	Студент демонстрирует опыт в области изучения на уровне указанных индикаторов.

Личностные качества	Студент демонстрирует ответственность в освоении результатов обучения на уровне запланированных индикаторов. Студент способен выносить суждения, делать оценки и формулировать выводы в области изучения. Студент может сообщать преподавателю и коллегам своего уровня собственное понимание и умения в области изучения.
---------------------	--

4.2. Для оценивания уровня выполнения критериев (уровня достижений обучающихся при проведении контрольно-оценочных мероприятий по дисциплине модуля) используется универсальная шкала (табл. 5).

Таблица 5

Шкала оценивания достижения результатов обучения (индикаторов) по уровням

Характеристика уровней достижения результатов обучения (индикаторов)				
№ п/п	Содержание уровня выполнения критерия оценивания результатов обучения (выполненное оценочное задание)	Шкала оценивания		
		Традиционная характеристика уровня		Качественная характеристика уровня
1.	Результаты обучения (индикаторы) достигнуты в полном объеме, замечаний нет	Отлично (80-100 баллов)	Зачтено	Высокий (В)
2.	Результаты обучения (индикаторы) в целом достигнуты, имеются замечания, которые не требуют обязательного устранения	Хорошо (60-79 баллов)		Средний (С)
3.	Результаты обучения (индикаторы) достигнуты не в полной мере, есть замечания	Удовлетворительно (40-59 баллов)		Пороговый (П)
4.	Освоение результатов обучения не соответствует индикаторам, имеются существенные ошибки и замечания, требуется доработка	Неудовлетворительно (менее 40 баллов)	Не зачтено	Недостаточный (Н)
5.	Результат обучения не достигнут, задание не выполнено	Недостаточно свидетельств для оценивания		Нет результата

5. СОДЕРЖАНИЕ КОНТРОЛЬНО-ОЦЕНОЧНЫХ МЕРОПРИЯТИЙ ПО ДИСЦИПЛИНЕ МОДУЛЯ

Задания по контрольно-оценочным мероприятиям в рамках текущей и промежуточной аттестации должны обеспечивать освоение и достижение результатов обучения (индикаторов) и предметного содержания дисциплины на соответствующем уровне.

5.1. Описание контрольно-оценочных мероприятий и средств текущего контроля по дисциплине модуля

5.1.1. Практические занятия

Не предусмотрено.

5.1.2. Лабораторные занятия

Номер занятия	Примерный перечень тем практических занятий
1	Text preprocessing for analysis.
2	Text vectorization.
3	Text classification using classical machine learning methods.
4	Text classification using deep neural networks.
5	The language model. Teaching a language model.
6	Automatic text generation.
7	Search for named entities in the text.
8	The mechanism of attention in neural networks. Networks with Transformer architecture.
9	Transfer of learning in word processing tasks.

5.1.3. Курсовая работа / Курсовой проект

Не предусмотрено.

5.1.4. Контрольная работа

Примерная тематика контрольных работ:

Pipeline design for natural language processing tasks.

Примерные задания в составе контрольных работ:

Design a sequence of actions to solve the problem of text analysis using machine learning. The pipeline must include:

1. The method of preparing the text for processing.
2. Approach to text tokenization.
3. Approach to text vectorization.
4. The machine learning model used.
5. The method of training the model.
6. The method of assessing the quality of the model.
7. Using a trained model to solve the problem of text analysis.
8. Other steps that may be needed when solving the problem.

Examples of natural language processing tasks for which you need to create pipelines:

- Text classification.
- Determination of the emotional coloring of the text.
- Automatic text generation.
- Search for named entities in the text.

5.1.5. Домашняя работа

Примерная тематика домашних работ:

Домашняя работа №1:

Training a language model for texts in Russian.

Домашняя работа №2:

Retraining of a pre-trained BERT network

Примерные задания в составе домашних работ:

1. Train a language model for the Russian language and use it to generate text. For performing the task, you need to do the following:

- Prepare a data set with texts in Russian. You can use ready-made datasets or create your own.
- Train the language model on the prepared data set.
- Using a trained language model, generate five examples of texts in Russian.
- Put the dataset, code, and trained model in open access on GitHub.
- Make a presentation or a technological article about the progress of the work, the justification of the decisions made and the results of the work.

* (Optional task). Record a video showing how the created solution works.

2. Train a pre-trained network with Transformer architecture for classifying texts in Russian. For performing the task, you need to do the following:

- Prepare a data set with texts in Russian for classification. You can use ready-made datasets or create your own.
- Select a pre-trained neural network with Transformer architecture suitable for the task of classifying texts in Russian.
- Perform additional training of the selected neural network on the prepared data set.
- Perform text classification testing using a trained neural network and evaluate the quality of the network.
- Put the dataset, code, and the completed model in open access on GitHub.
- Make a presentation or a technological article about the progress of work, justification of the decisions made and the results of work.

* (Optional task). Record a video showing how the created solution works.

An example of training a BERT neural network in TensorFlow –

https://www.tensorflow.org/text/tutorials/fine_tune_bert

Laptop with sample solution code –

https://colab.research.google.com/github/tensorflow/text/blob/master/docs/tutorials/fine_tune_bert.ipynb

Example of retraining neural networks with Transformer architecture in Hugging Face –

<https://huggingface.co/transformers/training.html>

5.1.6. Расчетная работа / Расчетно-графическая работа

Не предусмотрено.

5.1.7. Реферат / эссе / творческая работа

Не предусмотрено.

5.1.8. Проектная работа

Не предусмотрено.

5.1.9. Деловая (ролевая) игра / Дебаты / Дискуссия / Круглый стол
Не предусмотрено.

5.1.10. Кейс-анализ
Не предусмотрено.

5.2. Описание контрольно-оценочных мероприятий промежуточного контроля по дисциплине модуля

5.2.1. Зачет в форме независимого тестового контроля (НТК)
НТК по дисциплине модуля не проводится.

5.2.2. Зачет в традиционной форме (устные / письменные ответы на вопросы)

Список примерных вопросов для Зачета

1. Theoretical aspects of natural language processing.
2. Features of text processing in English.
3. Features of text processing in Russian.
4. Text preprocessing. Clearing the text. Deleting stop words/most and least frequent words.
5. Tokenization, stemming, lemmatization of the text.
6. Methods of text vectorization: building a dictionary, a bag of words.
7. Text vectorization methods: TF-IDF.
8. Text vectorization methods: word2vec.
9. Text vectorization methods: fasttext
10. Text vectorization methods: GloVe.
11. Classical machine learning methods for solving text classification problems.
12. Classical machine learning methods for solving the definition of the tonality of the text.
13. Neural network architectures for text processing: LSTM.
14. Neural network architectures for text processing: GRU.
15. Neural network architectures for text processing: one-dimensional convolutional networks.
16. Classification of text using neural networks.
17. Determination of the tonality of the text using neural networks.
18. Language model.
19. Teaching the language model.
20. Basic approaches to text generation.
21. The task of searching for named entities in the text.
22. The use of neural networks to search for named entities.
23. The mechanism of attention in neural networks.
24. Application of the attention mechanism for text processing.
25. Transformer neural network architecture.
26. Pre-trained neural networks for text processing BERT.
27. Pre-trained neural networks for GPT text processing.
28. Transfer of training for word processing tasks.
29. Classification of text using networks with Transformer architecture.
30. Text generation using networks with Transformer architecture.
31. Search for named entities in the text using networks with Transformer architecture.