Федеральное государственное образовательное бюджетное учреждение
высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Кафедра искусственного интеллекта
Факультет информационных технологий и анализа больших данных

Документ подписан усиленной неквалифицированной электронной подписью
Организация: Финансовый университет при Правительстве РФ
Утвержено: Проректор по учебной и методической работе Е.А. Каменева
Сертификат: O0wVO+GTEAYPxEfqbwibLr/bdppvpKZI
Дата: 23.09.2025 г.

С. В. Макрушин, В. А. Бартева

Современные технологии обработки текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
01.04.02 - Прикладная математика и информатика

Направленность программы магистратуры
Инженер машинного обучения

Рекомендовано
Факультет информационных технологий и анализа больших данных
(протокол №2 от 18.11.2025)

Одобрено
Кафедра искусственного интеллекта
(протокол №3 от 05.11.2025)

Москва 2025

Содержание

Наименование разделов РПД		**Стр.**
1.	Наименование дисциплины	5
2.	Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	5
3.	Место дисциплины в структуре образовательной программы	6
4.	Объем дисциплины(модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	6
5.	Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	7
5.1.	Содержание дисциплины	7
5.2.	Учебно-тематический план	8
5.3.	Содержание семинаров, практических занятий	9
6.	Перечень учебно-методического обеспечения для	12

	самостоятельной работы обучающихся по дисциплине	
6.1.	Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	12
6.2.	Перечень вопросов, заданий, тем для подготовки к текущему контролю	13
7.	Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	15
8.	Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	18
9.	Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	18
10.	Методические указания для обучающихся по освоению дисциплины	19
11.	Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных	20

	справочных систем	
12.	Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	21

1. Наименование дисциплины

«Современные технологии обработки текстов на естественных языках».

2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

Код компетенции	**Наименование компетенции**	**Индикатор достижения компетенции**	**Результаты обучения**
ПК-3	Способность разрабатывать прикладные решения с использованием методов анализа текстов, графов и сетевых структур	Использовать основные модели и методы обработки текстов на естественных языках, графов и семантических структур.	**Знать:** принципы и этапы обработки текстов на естественных языках (NLP): препроцессинг, токенизацию, векторизацию, нормализацию, синтаксический разбор, семантическое представление текста **Уметь:** реализовывать задачи NLP: классификация текстов, распознавание именованных объектов, анализ настроений, автоматическое суммирование, машинный перевод, диалоговые системы
		Решать с помощью	**Знать:** современные

		прикладных моделей NLP, графовой и сетевой аналитики прикладных задач предметной области	архитектуры моделей NLP: рекуррентные нейронные сети (RNN/LSTM), трансформеры (BERT, GPT), модели представлений слов (Word2Vec, GloVe) **Уметь:** создавать инструменты и программы для анализа больших объемов текстовых данных, включая разработку систем интеллектуального поиска, контент-анализа и автоматизированного анализа социальных медиа

3. Место дисциплины в структуре образовательной программы

Дисциплина «Современные технологии обработки текстов на естественных языках» относится к «Модулю дисциплин по выбору, углубляющих освоение программы магистратуры»

4. Объем дисциплины(модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

Очная форма обучения

Вид учебной работы	Всего (в з/е и часах)	Модуль 5 (в часах)

по дисциплине		
Общая трудоемкость дисциплины	**3/108**	**108**
Контактная работа - Аудиторные занятия	**32**	**32**
Лекции	8	8
Семинары, практические занятия	24	24
Самостоятельная работа	76	76
Вид текущего контроля	Контрольная работа;	Контрольная работа
Вид промежуточной аттестации	Зачет;	Зачет

5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема 1. Представление слов в виде векторов малой размерности

Представление слов в виде векторов малой размерности. Сингулярное разложение.

Концепция эмбеддинга. Роль эмбеддинга в глубоком обучении. Upstream и downstream задачи. Задачи self-supervising learning.

Алгоритм Word2vec, варианты Skip-Gram и CBoW. Алгоритм FastText . Предобученные модели эмбеннигов.

Тема 2. Рекуррентные нейронные сети в NLP

NLP Seq 2 seq задачи в глубоком обучении. Задача определения тональности текста. Машинный перевод. Задача выявления именованных сущностей и задача выделения частей речи.

Рекуррентные нейронные сети (RNN). Построение нейронных сетей на базе архитектуры RNN .

Рекуррентная ячейка LSTM . Рекуррентная ячейка GRU .

Модель ELMo .

Тема 3. Механизм внимания, BERT

Механизм внимания в NLP . Архитектура Transformer . Multi-head attention, cross-attention.

Модель Bidirectional Encoder Representations from Transformers (BERT). Предобученные реализации на базе BERT , дообучение моделей на базе BERT . Процедура дистилляции и модель DistillBERT . Модель RoBERTa .

Процедуры p re - training и f ine - tuning .

Тема 4. Большие лингвистические модели

Модель GPT (Generative Pre - Training). Авторегрессионный transformer decoder и особенности использования архитектуры задачи Transformer в задаче языкового моделирования.

Большие лингвистические модели (LLM). Виды больших языковых моделей. Особенности LLM , one short learning , few short learning . Промпт-инжениринг, техника цепочки рассуждений. Техника (Chain of Thought), Блокнот (Scratchpad).

Обзор современных LLM .

Тема 5. Технологии работы с LLM

Процедура обучения LLM . Proximal Policy Optimization для LLM.

Датасеты для обучения LLM

Методы оценки качества LLM .

Улучшенные методы обучения LLM , адаптеры. LoRA: Low-Rank Adaptation для LLM.

Квантизация для LLM.

Проблема галлюцинаций у LLM и архитектура Retrieval Augmented Generation (RAG).

5.2. Учебно-тематический план

Очная форма обучения

№ п/п	Наименование тем(разделов) дисциплины	Трудоемкость в часах				
		Всего	Контактная работа - Аудиторная работа			Самостоятельная работа
			Общая, в т.ч.:	Лекции	Семинары, практические занятия	
1	Представление слов в виде векторов малой размерности	22	6	2	4	16
2	Рекуррентные нейронные сети в NLP	22	7	2	5	15
3	Механизм внимания, BERT	22	7	2	5	15
4	Большие лингвистические модели	22	7	2	5	15
5	Технологии работы с LLM	20	5	0	5	15
	Итого	108	32	8	24	76

5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	**Перечень вопросов для обсуждения на семинарах, практических занятиях**	**Формы проведения занятий**

Представление слов в виде векторов малой размерности	1. Что такое эмбеддинг слова и какова его роль в глубоком обучении? 2. Как работает алгоритм Word2vec? В чем разница между Skip-Gram и CBoW? 3. Какие есть предобученные модели эмбеддингов и для каких задач их можно использовать? 4. Какое значение имеет сингулярное разложение при представлении слов в виде векторов?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Рекуррентные нейронные сети в NLP	1. В чем основное отличие между RNN, LSTM и GRU? 2. Какие задачи в NLP можно решать с помощью Seq2seq моделей на базе RNN? 3. Что такое модель ELMo и в чем ее преимущества перед классическими RNN? 4. Какие проблемы могут возникнуть при использовании RNN в NLP и как их можно решить?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Механизм внимания, BERT	1. Что такое механизм внимания и как он используется в архитектуре Transformer? 2. Как работает модель BERT и в чем ее основные преимущества перед другими архитектурами? 3. Что такое процедуры pre-training и fine-tuning в контексте работы с	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.

	моделями на базе BERT? 4. Какие существуют предобученные реализации на базе BERT и для каких задач они могут быть использованы?	
Большие лингвистические модели	1. Что такое Большие лингвистические модели (LLM) и в чем их основное преимущество? 2. Какие техники и методы используются для улучшения работы с LLM? 3. Что такое промпт-инжениринг и как он может быть применен в контексте LLM? 4. Какие проблемы могут возникнуть при работе с большими лингвистическими моделями и как их можно решить?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Технологии работы с LLM	1. Какие методы используются для оценки качества работы больших лингвистических моделей? 2. Что такое методы адаптеров и как они применяются в контексте LLM? 3. Что такое проблема галлюцинаций у LLM и как архитектура RAG помогает ее решить? 4. Какие техники квантизации могут быть применены для улучшения работы LLM?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
Представление слов в виде векторов малой размерности	Как происходит процесс обучения эмбеддингов, и какие факторы могут влиять на качество полученных векторов? Какие альтернативные методы представления слов в векторном пространстве существуют помимо Word 2 vec и FastText ?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Рекуррентные нейронные сети в NLP	Как влияет архитектура RNN на способность модели обучаться долгосрочным зависимостям в данных? Какие проблемы могут возникнуть при использовании RNN в задачах NLP , связанных с длинными последовательностям и?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Механизм внимания, BERT	Как механизм внимания улучшает производительность модели в сравнении с традиционными RNN или CNN ? В каких случаях использование	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.

	моделей на базе BERT может быть неэффективным или нежелательным?	
Большие лингвистические модели	Какие основные вызовы существуют при работе с большими лингвистическими моделями, такими как GPT или BERT ? Какие методы можно использовать для интерпретации работы больших лингвистических моделей и понимания их решений?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Технологии работы с LLM	Какие методы ансамблирования можно использовать для улучшения работы с большими лингвистическими моделями? В каких приложениях и сценариях использование LLM может привести к наилучшим результатам, и почему?	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерные вопросы контрольной работы

1. Рекуррентная нейронная сеть, принципы ее обучения

2. Модуль LSTM.

3. Модуль GRU.

4. Как механизм внимания улучшает работу моделей в задачах обработки текста?

5. Как BERT сравнивается с другими моделями в задачах NLP?

6. Как можно оптимизировать процесс fine-tuning моделей на базе BERT?

7. Какие вызовы возникают при работе с большими лингвистическими моделями и как их можно преодолеть?

8. В чем преимущества и недостатки использования больших лингвистических моделей для различных задач NLP?

9. Какие техники можно использовать для сокращения объема и ускорения работы больших лингвистических моделей?

10. В каких приложениях и сценариях использование адаптеров и квантизации может быть наиболее эффективным?

11. В чем отличие NLP от традиционной обработки текстов?

12. Какие методы используются для измерения качества языковых моделей?

13. В каких случаях предобученные модели эмбеддингов могут не быть эффективными?

14. Как влияет качество разметки на общую эффективность системы обработки текста?

15. Какие альтернативные методы разметки текста существуют помимо скрытых марковских моделей?

16. В чем отличие между LSTM и GRU, и какое из них лучше подходит для NLP?

17. Какие методы интерпретации работы больших лингвистических моделей вы знаете?

18. Какие методы адаптации и оптимизации существуют для улучшения работы больших лингвистических моделей?

19. Какие проблемы возникают при использовании тезаурусов в NLP?

20. Как рекуррентные нейронные сети решают проблему обработки последовательностей?

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических

рекомендациях Кафедры искусственного интеллекта Факультета информационных технологий и анализа больших данных.

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 2. *Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине.*

Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний

ПК-3 Способность разрабатывать прикладные решения с использованием методов анализа текстов, графов и сетевых структур

1) Использовать основные модели и методы обработки текстов на естественных языках, графов и семантических структур.

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: принципы и этапы обработки текстов на естественных языках (NLP): препроцессинг, токенизацию, векторизацию, нормализацию, синтаксический разбор, семантическое представление текста

Уметь: реализовывать задачи NLP: классификация текстов, распознавание именованных объектов, анализ настроений, автоматическое суммирование, машинный перевод, диалоговые системы

Типовые контрольные задания

Решить задачу классификации коротких текстов с помощью предобученной модели BERT.

2) Решать с помощью прикладных моделей NLP, графовой и сетевой аналитики прикладных задач предметной области

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: современные архитектуры моделей NLP: рекуррентные нейронные сети (RNN/LSTM), трансформеры (BERT, GPT), модели представлений слов (Word2Vec, GloVe)

Уметь: создавать инструменты и программы для анализа больших объемов текстовых данных, включая разработку систем интеллектуального поиска, контент-анализа и автоматизированного анализа социальных медиа

Типовые контрольные задания

Решить задачу определения тематики текста с помощью предобученной модели BERT.

Примеры практико-ориентированных заданий

1. Применить технологии обработки текста на естественном языке для выявления именованных сущностей из предложения.

2. Предложить наиболее подходящий алгоритм для определения тональности текста.

3. Дообучить языковую модель BERT для работы со специальным корпусом документов.

4. Создать прикладную модель идентификации некорректных текстовых высказываний.

5. Решить задачу определения тональности текстов из датасета texts . txt с помощью предобученной модели BERT.

Примерные вопросы для подготовки к Зачету

1. Структурные особенности текстов на естественном языке.

2. Извлечение именованных сущностей как задача разметки

3. Ранжированный информационный поиск

4. Вероятностный подход к парсингу

5. Синтаксис составляющих и синтаксис зависимостей

6. В чем разница между NLU и NLG?

7. Как НЛП связано с ИИ?

8. Различие подходов к решению задач NLP с помощью правил и с помощью алгоритмов машинного обучения.

9. Корпуса текстов, размеченные текстовые данные и их роль в решении задач NLP .

10. Назовите 5-7 задач NLP.

11. Каковы два основных типа анализа тональности?

12. Приведите пример акустической неоднозначности

13. Найти и аргументировать ошибку экспертной разметки 5 предложенных текстов (тексты заданы)

14. Составить схему разметки тональностей сообщений (по названиям заголовков)

15. В сети интернет найти 5 сообщений о перспективах запуска «Северного потока - 2» и представить выбранные новости в структуре «Название» - «Аннотация» - «Содержание» - «Оценка тональности». Обосновать выбранную меру оценки тональности.

16. Составить краткий аналитический обзор новостей с последнего лингвистического конгресса или симпозиума.

17. Опишите архитектуру модели CBOW в Word2vec и принципы ее обучения и применения.

18. Опишите архитектуру модели Skip-Gram.в Word2vec и принципы ее обучения и применения.

19. Задачи обработки текста: дистрибутивная семантика, матрица совместной встречаемости, представление слов в виде векторов малой размерности.

20. Рекуррентная нейронная сеть, принципы ее обучения. Сложности применения рекуррентных нейронных сетей.

21. Модуль LSTM.

22. Логика построения архитектуры ИНС BiLSTM .

23. Модуль GRU.

24. Механизм внимания в искусственных нейронных сетях.

25. Модель BERT. Принципы обучения модели BERT .

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

Основная литература:

1. Иванищева, О.Н. Прикладная лингвистика : Учебное пособие / О.Н. Иванищева Электрон. дан. Москва : Русайнс, 2024 235 с. Режим доступа: book.ru Internet access https://book.ru/book/953972 ISBN 978-5-466-06882-5. [БИК ID: RU\bookru\bibl\953972]

Дополнительная литература:

1. Шунейко, Александр Альфредович Квантитативная лингвистика и новые информационные технологии : учебник для вузов / А. А. Шунейко, И. А. Авдеенко. Электрон. дан. Москва : Юрайт, 2025 347 с (Высшее образование) URL: https://urait.ru/bcode/567628 (дата обращения: 01.09.2025). Режим доступа: Электронно-библиотечная система Юрайт, для авториз. пользователей https://urait.ru/bcode/567628 ISBN 978-5-534-15446-7 : 1719.00. [БИК ID: RU2fURAIT2f567628]

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

1. •Электронная библиотека Финансового университета (ЭБ) http://elib.fa.ru/

2. •Электронно-библиотечная система BOOK.RU http://www.book.ru

3. •Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» http://biblioclub.ru/

4. •Электронно-библиотечная система издательства «ЮРАЙТ» https://urait.ru/

5. •Электронно-библиотечная система издательства Проспект http://ebs.prospekt.org/books

6. •Электронно-библиотечная система издательства Лань https://e.lanbook.com/

7. •Деловая онлайн-библиотека Alpina Digital http://lib.alpinadigital.ru/

8. •Научная электронная библиотека eLibrary.ru http://elibrary.ru

9. •Национальная электронная библиотека http://нэб.рф/

10. Финансовая справочная система «Финансовый директор» http://www.1fd.ru/

11. http://www.aclweb.org/anthology/

12. http://www.nltk.org/book/

13. http://www.mitpressjournals.org/loi/coli

14. https://nlpub.ru

15. http://www.aot.ru/

10. Методические указания для обучающихся по освоению дисциплины

При изучении теоретического материала необходимо опираться на рабочую программу дисциплины, материалы лекций и литературу из основного списка. Кроме этого, необходимо активно работать с Интернет-источниками и пособиями других авторов, помогающими усвоить материал отдельных разделов программы.

Необходимо конспектировать лекции, помечая сложные и непонятные моменты с тем, чтобы задать вопросы лектору в конце лекции или же на консультации.

При подготовке к семинарским занятиям необходимо изучить вопросы, вынесенные на самостоятельное изучение, так как семинарские занятия предполагают их обсуждение и дискуссию по

теме; кроме того, задания для самостоятельной работы необходимы для того, чтобы успешно выполнить вынесенные самостоятельные задания на семинарах.

Индивидуальные задания для работы на компьютере, файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

Студентам при подготовке к контрольной работе следует использовать нормативные документы Финансового университета, Методические рекомендации по планированию и организации внеаудиторной самостоятельной работы студентов по образовательным программам бакалавриата и магистратуры в Финансовом университете, утвержденные приказом Финуниверситета от 11.05.2021 г. № 1040/о (см. сайт Финансового Университета: на главной странице раздел "Наш университет"→ "Единая правовая база Финуниверситета" → "Организация учебного процесса" → "Нормативные документы по самостоятельной работе").

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

- Комплект лицензионного программного обеспечения:

1. Kaspersky
2. Пакет офисных программ

- Современные профессиональные базы данных и информационные справочные системы:

1. Информационно-правовая система «Гарант»
2. Информационно-правовая система «Консультант Плюс»

3. Электронная энциклопедия: http://ru.wikipedia.org/wiki/Wiki

4. Система комплексного раскрытия информации «СКРИН» - http://www.skrin.ru/

5. Язык программирования Python 3. https://pythonworld.ru/

- Сертифицированные программные и аппаратные средства защиты информации

1. не предусмотрены

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

1. **Учебная аудитория** для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная оборудованием и техническими средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), набор демонстрационного оборудования (проектор, экран)

2. **Компьютерный класс** для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенный оборудованием и техническими средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), персональные компьютеры, набор демонстрационного оборудования (проектор, экран)