

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
РОССИЙСКОЙ ФЕДЕРАЦИИ»

Сертификат: gD9HS1NXFAVcyiy7fS3a3Jb7989RT4K0

Утверждено: Проректор по учебной и методической работе, Е.А. Каменева

Дата: 10.04.2026 г.

С.В. Макрушин , В.А. Бартева
Обработка текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки

09.03.03 - Прикладная информатика,

Образовательная программа «Прикладные информационные системы в экономике и
финансах»

Профиль:

«Прикладные информационные системы в экономике и финансах»

Рекомендовано

*Факультет информационных технологий и анализа больших данных
(протокол № 07 от 22.04.2026 г.)*

Одобрено

*Кафедра искусственного интеллекта
(протокол № 8 от 07.04.2026 г.)*



Содержание

1. Наименование дисциплины	3
2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	3
3. Место дисциплины в структуре образовательной программы	
4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	5
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	6
5.1. Содержание дисциплины	6
5.2. Учебно – тематический план	9
5.3. Содержание семинаров, практических занятий	10
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	13
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	13
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю	16
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	17
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	7
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	15
10. Методические указания для обучающихся по освоению дисциплины	25
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)	27
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	27



1. Наименование дисциплины

Обработка 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	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКП-1	Способен разрабатывать и отлаживать программный код	Выполняет формализацию и алгоритмизацию поставленных задач для разработки программного кода	Знать: методы формализации и алгоритмического представления задач для последующей реализации программного кода Уметь: формализовать поставленные задачи и создавать эффективные алгоритмы для последующего написания качественного программного кода
		Оформляет программный код в соответствии с установленными требованиями	Знать: правила оформления программного кода согласно принятым стандартам и соглашениям Уметь: оформлять программный код грамотно и структурированно, соблюдая требования читаемости, документирования и согласованности с корпоративными стандартами качества



Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКП-6	Способен участвовать в промышленной разработке программного обеспечения	Работает в соответствии с промышленными методологиями разработки	Знать: основные промышленные методологии разработки ПО и принципы их эффективного применения Уметь: применять выбранную промышленную методологию разработки в повседневной работе над проектами, взаимодействуя с командой и обеспечивая соблюдение стандартов качества и сроков исполнения
		Применяет инструменты поддержки методологий разработки и использует навыки ведения документации в соответствии с требованиями методологии	Знать: типовые инструменты управления проектами и системами контроля версий Уметь: использовать инструменты поддержки методологий разработки и вести техническую документацию в строгом соответствии с установленными требованиями выбранной методологии



3. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка текстов на естественных языках» относится к «Модулю "Методы искусственного интеллекта"».

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

очно-заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 8 (в часах)
Общая трудоемкость дисциплины	3/108	108
Контактная работа – Аудиторные занятия	34	34
<i>Лекции</i>	<i>16</i>	<i>16</i>
<i>Семинары, практические занятия</i>	<i>18</i>	<i>18</i>
<i>Самостоятельная работа</i>	<i>74</i>	<i>74</i>
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. Введение в NLP, базовая обработка текста и дистанция редактирования

Особенности текстов на естественном языке. Неоднозначность на всех уровнях языка. Основные задачи автоматического анализа текстов. Исторический обзор и развитие NLP.

Основные подходы к решению задач: правила, написанные вручную и машинное обучение. Основные этапы обработки текста. Обзор инструментов и библиотек для NLP.

Предобработка текста: токенизация и сегментация.

Метрики расстояния между строками и между словами. Расстояние Левенштейна. Алгоритм Вагнера – Фишера.

Нормализация слов: стемминг, лемматизация, морфологические анализаторы.

Корпуса текстов. Русскоязычные корпуса текстов.

Закон Ципфа. Закон Стоп-слова.

Тема 2. Языковые модели и векторная семантика

Семантика в лингвистике. Семантические роли и отношения между словами. Тезаурусы, WordNet.

N-граммы. Перплексия. Методы сглаживания, линейная интерполяция.

Применение языковых моделей: предсказание ввода, исправление ошибок правописания, распознавание речи, порождение текста.

Дистрибутивная гипотеза. Векторная семантика. Метрики в векторных пространствах, косинусная мера сходства.

Метрика совместной встречаемости. Мешок слов. Векторное представление документа.

Тема 3. Представление слов в виде векторов малой размерности

Представление слов в виде векторов малой размерности. Сингулярное разложение.

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



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

Тема 4. Задачи разметки текста и скрытые марковские модели

Разметка по частям речи.

Извлечение именованных сущностей как задача разметки.

Скрытые марковские модели, их достоинства и недостатки.

Модификации скрытых марковских моделей.

Тема 5. Рекуррентные нейронные сети в NLP

NLP Seq2seq 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 6. 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 pre-training и fine-tuning.

Тема 7. 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 shot learning, few shot learning. Промпт-инжиниринг, техника цепочки рассуждений. Техника (Chain of Thought), Блокнот (Scratchpad).

Обзор современных LLM.

Тема 8. 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

Очно-заочная 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	Введение в NLP, базовая обработка текста и дистанция редактирования	31	6	2	4	25
2	Языковые модели и векторная семантика	11	4	2	2	7
3	Представление слов в виде векторов малой размерности	11	4	2	2	7
4	Задачи разметки текста и скрытые марковские модели	11	4	2	2	7
5	Рекуррентные нейронные сети в NLP	11	4	2	2	7
6	Механизм внимания, BERT	11	4	2	2	7
7	Большие лингвистические модели	11	4	2	2	7
8	Технологии работы с LLM	11	4	2	2	7
В целом по дисциплине		108	34	16	18	74



5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
Введение в NLP, базовая обработка текста и дистанция редактирования	Какие основные задачи решает NLP и какие этапы обработки текста вы можете выделить? В чем разница между стеммингом и лемматизацией? Какой из этих методов вы бы использовали для русскоязычного текста и почему? Что такое расстояние Левенштейна и в каких ситуациях оно полезно? Что означает Закон Ципфа в контексте NLP и как он применяется?	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Языковые модели и векторная семантика	Что такое N-граммы и какая их роль в языковых моделях? Какие метрики используются для измерения семантической близости в векторных пространствах? Чем отличаются методы Skip-Gram и CBoW в контексте Word2vec? Какова роль у WordNet в языковых моделях и в чем его особенности?	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Представление слов в виде векторов малой размерности	Что такое эмбединг слова и какова его роль в глубоком обучении? Как работает алгоритм Word2vec? В чем разница между Skip-Gram и CBoW? Какие есть предобученные модели эмбедингов и для каких задач их можно использовать? 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.
Задачи разметки текста и скрытые марковские модели	Что такое разметка по частям речи и зачем она нужна в NLP? В чем основные преимущества и недостатки	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и



Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
	скрытых марковских моделей для разметки текста? Какие модификации скрытых марковских моделей вы знаете и для каких задач они могут быть полезны? Как извлечение именованных сущностей связано с задачей разметки текста?	коллективное обсуждение решений.
Рекуррентные нейронные сети в NLP	В чем основное отличие между RNN, LSTM и GRU? Какие задачи в NLP можно решать с помощью Seq2seq моделей на базе RNN? Что такое модель ELMo и в чем ее преимущества перед классическими RNN? 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	Что такое механизм внимания и как он используется в архитектуре Transformer? Как работает модель BERT и в чем ее основные преимущества перед другими архитектурами? Что такое процедуры pre-training и fine-tuning в контексте работы с моделями на базе BERT? Какие существуют предобученные реализации на базе BERT и для каких задач они могут быть использованы?	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Большие лингвистические модели	Что такое Большие лингвистические модели (LLM) и в чем их основное преимущество? Какие техники и методы используются для улучшения работы с LLM? Что такое промпт-инжиниринг и как он может быть применен в контексте LLM? Какие проблемы могут возникнуть при	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.



Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
	работе с большими лингвистическими моделями и как их можно решить?	
Технологии работы с LLM	Какие методы используются для оценки качества работы больших лингвистических моделей? Что такое методы адаптеров и как они применяются в контексте LLM? Что такое проблема галлюцинаций у LLM и как архитектура RAG помогает ее решить? 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
Введение в NLP, базовая обработка текста и дистанция редактирования	Как влияют стоп-слова на качество анализа текста, и какие методы их идентификации существуют? Какие основные этапы предварительной обработки текста вы бы выделили, и почему каждый из них важен?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Языковые модели и векторная семантика	Как векторное представление документа отличается от векторного представления слова, и в каких задачах оно применяется? Какие основные методы сглаживания N-грамм вы знаете и как они влияют на качество языковой модели?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Представление слов в виде векторов малой размерности	Как происходит процесс обучения эмбедингов, и какие факторы могут влиять на качество полученных векторов? Какие альтернативные методы представления слов в векторном пространстве существуют помимо Word2vec и FastText?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Задачи разметки текста и скрытые марковские модели	Какие проблемы могут возникнуть при разметке текста, и как их можно минимизировать с помощью скрытых марковских моделей? В чем особенности применения скрытых марковских моделей для извлечения информации из текста?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.



Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
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 последовательностями?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Механизм внимания, BERT	Как механизм внимания улучшает производительность модели в сравнении с традиционными RNN или CNN? В каких случаях использование моделей на базе BERT может быть неэффективным или нежелательным?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
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?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
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?	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.



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

Примерные вопросы для подготовки к контрольной работе

1. Какие основные вызовы и проблемы существуют в обработке текстов на естественном языке?
2. Какие факторы влияют на эффективность моделей NLP?
3. Какие этапы обработки текста важны для достижения высокого качества анализа?
4. Как можно применить машинное обучение для автоматического анализа текстов?
5. Что такое баланс между правилами, написанными вручную, и методами машинного обучения в NLP?
6. Как векторная семантика помогает в решении задач обработки текста?
7. Каковы преимущества использования эмбедингов перед традиционными методами представления текста?
8. Как можно адаптировать предобученные эмбединги для конкретной задачи?
9. В чем основные преимущества использования скрытых марковских моделей для разметки текста?
10. Какие проблемы могут возникнуть при использовании RNN и как их можно решить?
11. Как механизм внимания улучшает работу моделей в задачах обработки текста?
12. Как BERT сравнивается с другими моделями в задачах NLP?
13. Как можно оптимизировать процесс fine-tuning моделей на базе BERT?
14. Какие вызовы возникают при работе с большими лингвистическими моделями и как их можно преодолеть?
15. В чем преимущества и недостатки использования больших лингвистических моделей для различных задач NLP?
16. Какие техники можно использовать для сокращения объема и ускорения работы больших лингвистических моделей?
17. В каких приложениях и сценариях использование адаптеров и квантизации может быть наиболее эффективным?
18. В чем отличие NLP от традиционной обработки текстов?
19. Какие методы используются для измерения качества языковых моделей?
20. В каких случаях предобученные модели эмбедингов могут не быть эффективными?
21. Как влияет качество разметки на общую эффективность системы обработки текста?



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

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

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

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

26. Какие проблемы возникают при использовании тезаурусов в NLP?

27. Как рекуррентные нейронные сети решают проблему обработки последовательностей?

Дополнительная информация

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

Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания
ПКП-1. Способен разрабатывать и отлаживать программный код	Выполняет формализацию и алгоритмизацию поставленных задач для разработки программного кода	Знать: методы формализации и алгоритмического представления задач для последующей реализации программного кода Уметь: формализовать поставленные задачи и создавать эффективные алгоритмы для последующего написания каче-	Используйте алгоритмы кластеризации, например, на основе K-means, для группировки текстов по схожести содержания. Реализуйте алгоритм кластеризации и оцените качество его работы с использованием релевантных метрик.



Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания
		ственного программного кода	
	Оформляет программный код в соответствии с установленными требованиями	Знать: правила оформления программного кода согласно принятым стандартам и соглашениям Уметь: оформлять программный код грамотно и структурированно, соблюдая требования читаемости, документирования и согласованности с корпоративными стандартами качества	Используйте методы анализа сентимента для оценки влияния экономических новостей и сообщений на рыночные индексы или активы. Анализируйте эмоциональную окраску новостных статей и их влияние на изменение цен и торговые объемы.
ПКП-6. Способен участвовать в про-	Работает в соответствии с промышлен-	Знать: основные промышленные методологии разработки	Разработайте прототип чат-бота с функционалом обработки текста на естественном языке для предоставления финансовых консультаций или информации. Используйте



Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания
мышленной разработке программного обеспечения	ными методологиями разработки	ПО и принципы их эффективного применения Уметь: применять выбранную промышленную методологию разработки в повседневной работе над проектами, взаимодействуя с командой и обеспечивая соблюдение стандартов качества и сроков исполнения	современные библиотеки и фреймворки для реализации функционала NLP и проведите тестирование прототипа, оценив его эффективность и точность ответов.
	Применяет инструменты поддержки методологий разработки и использует навыки ведения документации в соответствии с	Знать: типовые инструменты управления проектами и системами контроля версий Уметь: использовать инструменты под-	Какие основные методы препроцессинга текста вы бы использовали перед анализом? Приведите примеры нескольких методов и объясните, для каких задач они наиболее эффективны.



Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания
	требованиями методологии	держки методологий разработки и вести техническую документацию в строгом соответствии с установленными требованиями выбранной методологии	



Примеры практико-ориентированных заданий

Задание 1. Проведите частотный анализ лемматизированного корпуса: составьте топ-20 наиболее частотных токенов (после удаления стоп-слов) и таблицу с их абсолютной и относительной частотой. Оцените, соответствует ли распределение частот закону Ципфа: постройте описание ожидаемой кривой ранг–частота и сравните с реальными данными выборки. Сделайте вывод о применимости закона к данному типу текстов.

Задание 2. Представьте три документа (обращения) в виде TF-IDF векторов. Документы: D1 = «Требую выплату по страховому случаю КАСКО, произошедшему 15 февраля»; D2 = «Прошу вернуть страховую премию при досрочном расторжении договора ОСАГО»; D3 = «Прошу произвести перерасчёт страховой премии по договору страхования имущества». Эталонные категории риска: R1 = «страховая выплата», R2 = «возврат премии», R3 = «изменение условий». Рассчитайте косинусное сходство каждого документа с каждой категорией риска (используйте ключевые слова категорий как псевдодокументы) и отнесите каждое обращение к наиболее близкой категории.

Задание 3. Выполните ручную NER-разметку следующего новостного текста в формате BIO-тегирования. Используйте теги: B-ORG/I-ORG (организация), B-PER/I-PER (персона), B-LOC/I-LOC (место), B-DATE/I-DATE (дата), B-MONEY/I-MONEY (денежная сумма), B-PCT/I-PCT (процент), O (вне сущности). ТЕКСТ: Совет директоров ПАО Газпром 28 февраля принял решение о выплате дивидендов в размере 51,03 рубля на акцию. Председатель правления Алексей Миллер сообщил, что выручка компании за 2023 год составила 8,5 трлн рублей. Аналитики Сбербанк Инвестиции прогнозируют рост котировок на Московской бирже на 12% в первом квартале. Представьте результат в виде таблицы Токен - Тег - Тип сущности.

Примеры вопросов для подготовки к промежуточной аттестации

Примерные вопросы для подготовки к зачету

1. NLP как одна из ведущих областей искусственного интеллекта.
2. Естественный язык как объект автоматической обработки.
3. Популярные задачи NLP и общие подходы к их решению.
4. Предварительная обработка текста. Регулярные выражения.
5. Стеммеры, лемматизаторы, морфологические анализаторы.
6. N-граммы. Дистрибутивная гипотеза. Матрица совместной встречаемости.



7. Применение языковых моделей: предсказание ввода, исправление ошибок правописания.
8. Проблемы с языковыми моделями и их решения.
9. Проблемы с тегами; полезность автоматических аннотаций.
10. Скрытые марковские модели, их плюсы и минусы.
11. Классификация текстов: постановка задачи и методы.
12. Наивный байесовский классификатор. Проблемы с классификацией текста.
13. Анализ тональности, извлечение аспектов
14. Меры оценки системы NLP.
15. Поиск информации. Бинарный поиск.
16. Поиск информации. Фразовые запросы.
17. TF-IDF.
18. Лексические базы данных. WordNet – организация, специфика, применение.
19. Семантическое сходство. Недистрибутивные методы.
20. Семантическое сходство. Фразовые запросы. Косинусное расстояние / подобие.
21. Python как язык программирования и инструмент для написания проектов NLP.
22. Векторная модель word2vec.
23. Векторная модель BERT.
24. Машинное обучение в НЛП.
25. Обзор RapidMiner и Orange.



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

Основная литература:

1. Махлина, Светлана Тевельевна Семиотика культуры повседневности : учебник для вузов / С. Т. Махлина. 2-е изд. Электрон. дан. Москва : Юрайт, 2025 255 с (Высшее образование) URL: <https://urait.ru/bcode/567559> (дата обращения: 01.09.2025). Режим доступа: Электронно-библиотечная система Юрайт, для авториз. пользователей <https://urait.ru/bcode/567559> ISBN 978-5-534-13763-7 : 1319.00. [БИК ID: RU2fURAIT2f567559]
2. 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. Влавацкая, Марина Витальевна Введение в языкознание : Учебное пособие / Новосибирский государственный технический университет Новосибирск : Новосибирский государственный технический университет (НГТУ), 2019 416 с. ВО - Бакалавриат <https://znanium.com/catalog/document?id=397618> ISBN 978-5-7782-3727-8. [БИК ID: RU\infra-m\znanium\bibl\1869260]

Нормативные правовые акты:

-

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

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/> (<http://library.fa.ru/files/elibfa.pdf>)
2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
3. Электронно-библиотечная система "Университетская библиотека ОНЛАЙН" <http://biblioclub.ru/>
4. "Деловая онлайн библиотека" издательства "Альпина Паблишер" <http://lib.alpinadigital.ru/en/library>
5. Электронно-библиотечная система издательства "Лань" <https://e.lanbook.com/>
6. Образовательная платформа "ЮРАЙТ" <https://urait.ru/>
7. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>



8. • Электронно-библиотечная система издательства Проспект
<http://ebs.prospekt.org/books>
9. • Национальная электронная библиотека <http://нэб.рф/>
10. Финансовая справочная система «Финансовый директор» <http://www.1fd.ru/>



10. Методические указания для обучающихся по освоению дисциплины

Основные этапы работы студента по дисциплине **Обработка текстов на естественных языках**:

1. Предварительная ориентировка в подлежащем изучению учебном материале по программе.
2. Ознакомление с рекомендованной учебной литературой.
3. Слушание и конспектирование лекций, а также выполнение других видов учебной работы.
4. Планирование самостоятельной работы.
5. Обобщение и систематизация информации, почерпнутой из лекций и прочитанной литературы.

6. Выполнение контрольной работы.

Рекомендации по работе с учебным материалом:

1. Осознавайте наличный уровень полученных вами знаний.
2. В ситуации непонимания нужно выявить тот первичный уровень и факторы непонимания, которые стали препятствием понимания последующего.
3. Задавайте сами себе вопросы и пытайтесь ответить на них.

Рекомендации по работе на лекции и с лекционным материалом:

1. Основная задача на лекции – осмысление излагаемого в ней материала. Для этого необходимо слушать лекцию с самого начала, не упуская общих, ориентирующих в материале рассуждений и установок лектора.
2. Ведение записей на лекции важно и полезно для лучшего осмысливания материала, для сохранения информации, с целью ее дальнейшего использования.
3. Для облегчения записи рекомендуется применять сокращения повторяющихся терминов или хорошо известных понятий.

Рекомендации по работе с литературой:

1. Если возникли затруднения при разыскивании материала, по какому-либо конкретному вопросу, следует обратиться к предметному указателю, напечатанному, как правило, в конце каждого литературного источника.
2. Предметный указатель – это алфавитный список основных научных понятий (терминов), содержание которых раскрыто в книге, рядом с термином стоят числа, обозначающие номера страниц, на которых изложен материал, относящийся к данному понятию.

Рекомендации по выполнению контрольной работы:



1. Перед выполнением контрольной работы студент должен изучить соответствующие разделы учебной литературы.

2. Контрольную работу студент должен выполнять самостоятельно, используя те навыки и умения, которые получил на лекциях и практических занятиях.

3. При затруднениях, возникших при выполнении контрольной работы, студент может получить консультацию преподавателя.

Студентам при подготовке контрольной работы 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. Python 3.8
2. Jupyter Notebook (Windows 10)
3. Антивирус Kaspersky

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)