

**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

Сертификат: /rjxVXXdGF7bzxxAgzbtrTry8yQuR2u6

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

Дата: 27.01.2026 г.

С.В. Макрушин, В.А. Бартева

Обработка текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

09.03.04 - Программная инженерия,

Образовательная программа «Инженер-разработчик программного обеспечения»

Профиль:

«Инженер-разработчик программного обеспечения»

Рекомендовано

*Факультет информационных технологий и анализа больших данных
(протокол № 06 от 17.03.2026 г.)*

Одобрено

*Кафедра искусственного интеллекта
(протокол № 7 от 03.03.2026 г.)*

Содержание

1. Наименование дисциплины	4
2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	4
3. Место дисциплины в структуре образовательной программы	7
4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	7
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий . . .	8
5.1. Содержание дисциплины	8
5.2. Учебно-тематический план	11
5.3. Содержание семинаров	12
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	15
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	15
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю . . .	17
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	19
8. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины	24
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	25
10. Методические указания для обучающихся по освоению дисциплины	26
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	28

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	29
--	----

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	
ПКП-1	Способность описывать, анализировать и проектировать интерфейс программных модулей с учетом требований к ним	
Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	
Демонстрирует знания основных понятий интерфейсов программных модулей, понятие внешней и внутренней среды, читает и понимает готовую программную документацию в части описания интерфейсов	знать: интерфейсы модулей библиотек для решения задач NLP уметь: использовать библиотеки для решения задач NLP	
Понимает достоинства и недостатки различных архитектурных решений в области проектирования интерфейсов программных модулей, может критически анализировать существующие решения	знать: архитектуры ИНС для решения задач NLP уметь: создавать ИНС на архитектурах типичных для решения задач NLP	

Описывает интерфейс программной системы в формализованном виде по определенным стандартам, демонстрирует знания общепринятых стандартов описания архитектуры программной системы	знать: способы формального описания архитектур ИНС, типичных для решения задач NLP уметь: описывать архитектуры реальных ИНС, использующихся для решения задач ИНС
Проектирует интерфейс программного модуля с учетом требований к программной системе в целом и с учетом интеграции с другими программными модулями	знать: подходы к проектированию архитектур ИНС для решения задачи NLP уметь: проектировать архитектуру ИНС для решения задачи NLP
ОПК-3	Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности
Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
Проводит самостоятельный поиск информации в открытых источниках по определенной заданной тематике	знать: источники информации о современных моделях для решения задач NLP уметь: искать современные модели для решения распространенных задач NLP
Проводит систематический обзор источников информации, анализировать содержащиеся в них данные, делать и обосновывать выводы на основе проведенного обзора	знать: источники информации о моделях NLP и подходы к проведению их анализа уметь: делать выводы о наиболее эффективных подходах к решению задач NLP

Демонстрирует знания основных требований информационной безопасности, основных алгоритмов защиты информации, в том числе с использованием криптографических протоколов	знать: подходы к атакам на модели NLP уметь: предотвращать атаки на модели NLP
---	--

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 трудоёмкость дисциплины	3/108	108
Контактная работа – Аудиторные занятия	26	26
Лекции	8	8
Семинары, практические занятия	18	18
Самостоятельная работа	82	82
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, oneshortlearning, fewshotlearning. Промпт-инжиниринг, техника цепочки рассуждений. Техника (ChainofThought), 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

№ п/п	Наименование тем(разделов) дисциплины	Трудоемкость в часах				
		Всего	Контактная работа - Аудиторная работа			Самостоя- тельная работа
			Общая, в т.ч.:	Лекции	Семинары, практическ- ие занятия	
1	Введение в NLP, базовая обработка текста и дистанция редактирования	15	4	2	2	11
2	Языковые модели и векторная семантика	14	4	2	2	10
3	Представление слов в виде векторов малой размерности	14	4	2	2	10
4	Задачи разметки текста и скрытые марковские модели	15	4	2	2	11
5	Рекуррентные нейронные сети в NLP	12	2	0	2	10
6	Механизм внимания, BERT	12	2	0	2	10
7	Большие лингвистические модели	12	2	0	2	10
8	Технологии работы с LLM	14	4	0	4	10
	Итого	108	26	8	18	82

Формы текущего контроля успеваемости согласно учебному плану: **Контрольная работа.**

5.3. Содержание семинаров

Наименование тем (разделов) дисциплины	
Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
Тема 1. Введение в NLP, базовая обработка текста и дистанция редактирования	
1. Какие основные задачи решает NLP и какие этапы обработки текста вы можете выделить? 2. В чем разница между стеммингом и лемматизацией? Какой из этих методов вы бы использовали для русскоязычного текста и почему? 3. Что такое расстояние Левенштейна и в каких ситуациях оно полезно? 4. 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.
Тема 2. Языковые модели и векторная семантика	
1. Что такое N-граммы и какая их роль в языковых моделях? 2. Какие метрики используются для измерения семантической близости в векторных пространствах? 3. Чем отличаются методы Skip-Gram и CBoW в контексте Word2vec? 4. Какая 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.
Тема 3. Представление слов в виде векторов малой размерности	
1. Что такое эмбединг 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 эмбедингов 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.

Тема 4. Задачи разметки текста и скрытые марковские модели	
1. Что такое разметка по частям речи и зачем она нужна в NLP? 2. В чем основные преимущества и недостатки скрытых марковских моделей для разметки текста? 3. Какие модификации скрытых марковских моделей вы знаете и для каких задач они могут быть полезны? 4. Как извлечение именованных сущностей связано с задачей разметки текста?	Интерактивная форма. Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений.
Тема 5. 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.
Тема 6. 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 моделями на базе BERT? 4. 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.
Тема 7. 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-инжиниринг 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.

Тема 8. 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
Тема 1. 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.
Тема 2. 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.
Тема 3. 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.

Тема 4. 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.
Тема 5. 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.
Тема 6. 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.
Тема 7. 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.
Тема 8. 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 контрольной работы

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?

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

ПКП-1 Способность описывать, анализировать и проектировать интерфейс программных модулей с учетом требований к ним

1) Демонстрирует знания основных понятий интерфейсов программных модулей, понятие внешней и внутренней среды, читает и понимает готовую программную документацию в части описания интерфейсов

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

Знать: интерфейсы модулей библиотек для решения задач NLP

Уметь: использовать библиотеки для решения задач NLP

Типовые контрольные задания

Решите задачу токенизации текста и удаления стоп слов с помощью известных библиотек NLP.

2) Понимает достоинства и недостатки различных архитектурных решений в области проектирования интерфейсов программных модулей, может критически анализировать существующие решения

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

Знать: архитектуры ИНС для решения задач NLP

Уметь: создавать ИНС на архитектурах типичных для решения задач NLP

Типовые контрольные задания

Постройте архитектуру ИНС для решения задачи классификации коротких текстов по тональности (положительная / нейтральная / отрицательная).

3) Описывает интерфейс программной системы в формализованном виде по определенным стандартам, демонстрирует знания общепринятых стандартов описания архитектуры программной системы

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

Знать: способы формального описания архитектур ИНС, типичных для решения задач NLP

Уметь: описывать архитектуры реальных ИНС, использующихся для решения задач ИНС

Типовые контрольные задания

Опишите модель ИНС для решения задач NLP, включая указания размерности тензоров передаваемых между слоями модели.

4) Проектирует интерфейс программного модуля с учетом требований к программной системе в целом и с учетом интеграции с другими программными модулями

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

Знать: подходы к проектированию архитектур ИНС для решения задачи NLP

Уметь: проектировать архитектуру ИНС для решения задачи NLP

Типовые контрольные задания

Спроектируйте многослойную нейронную сеть различающую язык твита написанного в кодировке Unicode.

ОПК-3 Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности

1) Проводит самостоятельный поиск информации в открытых источниках по определенной заданной тематике

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

Знать: источники информации о современных моделях для решения задач NLP

Уметь: искать современные модели для решения распространенных задач NLP

Типовые контрольные задания

Найдите современные модели позволяющие провести классификацию тональности текста.

2) Проводит систематический обзор источников информации, анализировать содержащиеся в них данные, делать и обосновывать выводы на основе проведенного обзора

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

Знать: источники информации о моделях NLP и подходы к проведению их анализа

Уметь: делать выводы о наиболее эффективных подходах к решению задач NLP

Типовые контрольные задания

Выберите одну из современных моделей позволяющую провести классификацию тональности текста. Обоснуйте свой выбор.

3) Демонстрирует знания основных требований информационной безопасности, основных алгоритмов защиты информации, в том числе с использованием криптографических протоколов

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

Знать: подходы к атакам на модели NLP

Уметь: предотвращать атаки на модели NLP

Типовые контрольные задания

Проведите тестирования алгоритма машинного обучения на переобучение.

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.

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 литературы, необходимой для освоения дисциплины

Основная 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]
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]
3. Махлина, Светлана Тевельевна Лингвистика и семиотика : учебник и практикум для вузов / С. Т. Махлина. Электрон. дан. Москва : Юрайт, 2026 260 с (Высшее образование) URL: <https://urait.ru/bcode/588682> (дата обращения: 22.01.2026). Режим доступа: Электронно-библиотечная система Юрайт, для авториз. пользователей <https://urait.ru/bcode/588682> ISBN 978-5-534-14194-8 : 1339.00. [БИК ID: RU2fURAIT2f588682]

Дополнительная литература:

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 Пабли-
шер" <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.пф/>
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. Пакет офисных программ
4. 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/>

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)