

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

Кафедра искусственного интеллекта
Факультета информационных технологий и анализа больших данных

УТВЕРЖДАЮ

Проректор по учебной
и методической работе

_____ Е.А. Каменева

11.11.2024 г.

С.В. Макрушин, В.А. Малекова

Обработка текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

09.03.03-Прикладная информатика,

ОП «Прикладные информационные системы в экономике и финансах»

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол № 48 от 30.10.2024 г.)*

*Одобрено заседанием Кафедры искусственного интеллекта
(протокол № 2 от 01.10.2024 г.)*

Москва 2024

СОДЕРЖАНИЕ

1.Наименование дисциплины.....	2
2.Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	2
3.Место дисциплины в структуре образовательной программы.....	3
4.Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся (в семестре, в сессию)	3
5.Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий.....	4
5.1. Содержание дисциплины	4
5.2. Учебно-тематический план	6
5.3. Содержание семинаров, практических занятий.....	8
6. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине.....	10
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	10
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю .	12
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине.....	14
8.Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	16
9.Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	17
10.Методические указания для обучающихся по освоению дисциплины.	17
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	19
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	20

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-6	Способен разрабатывать, реализовывать и применять методы интеллектуального анализа данных и машинного обучения для автоматизации решения неструктурированных и слабоструктурированных задач экономических предметных областей	Использует знания современных методов интеллектуального анализа данных (в том числе, больших данных) и способы их программной реализации	Знать: подходы к выполнению аналитических исследований в области обработки больших данных на естественном языке. Уметь: выполнять аналитические исследования в области обработки больших данных на естественном языке.
		Осуществляет поиск, сбор, анализ и интерпретацию данных экономических предметных областей с применением методов искусственного интеллекта и машинного обучения	Знать: методы использования текстовых данных для задач в экономических областях с использованием технологий обработки данных на естественном языке. Уметь: применять методы использования текстовых данных для задач в экономических областях на основе технологий обработки данных на естественном языке.
		Владеет современным инструментарием искусственного интеллекта и его использованием при разработке и развитии существующих финансово-экономических информационных систем	Знать: ключевые фреймворки и библиотеки для решения задач обработки текста на естественном языке. Уметь: строить модели по обработке текста на естественном языке с использованием современных фреймворков и библиотек и использовать их для развития существующих финансово-экономических информационных систем.

3. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка текстов на естественных языках» относится к Циклу профиля (элективный) по направлению подготовки 09.03.03 - Прикладная информатика, ОП «Прикладные информационные системы в экономике и финансах».

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

очная форма обучения / очно-заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 7 / 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>
Самостоятельная работа	74	74
Вид текущего контроля	Контрольная работа	Контрольная работа
Вид промежуточной аттестации	зачет	зачет

заочная форма обучения (ИОО)

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	10	10
<i>Лекции</i>	<i>2</i>	<i>2</i>
<i>Семинары, практические занятия</i>	<i>8</i>	<i>8</i>
Самостоятельная работа	98	98
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

Тема1. 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

очная форма обучения, очно-заочная форма обучения

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости
		Всего	*Контактная работа - Аудиторная работа			Самостояте льная работа	
			Общая, в т.ч.:	Лекции	Семинары, практическ ие занятия		
1.	Введение в NLP, базовая обработка текста и дистанция редактирования	31	6	2	4	25	Самостоятельн ые работы. Участие в решении задач на практических занятиях. Собеседования
2.	Языковые модели и векторная семантика	11	4	2	2	7	

3.	Представление слов в виде векторов малой размерности	11	4	2	2	7	по домашним заданиям. Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.
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	Согласно учебному плану: контрольная работа
	В целом по дисциплине	108	34	16	18	74	
	Итого в %		31	47	53	69	

заочная форма обучения (ИОО)

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости
		Всего	*Контактная работа - Аудиторная работа			Самостояте льная работа	
			Общая, в т.ч.:	Лекции	Семинары, практическ ие занятия		
1.	Введение в NLP, базовая обработка текста и дистанция редактирования	7	3	2	1	4	Самостоятельн ые работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.
2.	Языковые модели и векторная семантика	13	1	0	1	12	
3.	Представление слов в виде векторов малой размерности	14	1	0	1	13	
4.	Задачи разметки текста и скрытые марковские модели	13	1	0	1	12	

5.	Рекуррентные нейронные сети в NLP	15	1	0	1	14	Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.
6.	Механизм внимания, BERT	14	1	0	1	13	
7.	Большие лингвистические модели	16	1	0	1	15	
8.	Технологии работы с LLM	16	1	0	1	15	
	В целом по дисциплине	108	10	2	8	98	Согласно учебному плану: контрольная работа
	Итого в %		9	20	80	91	

* объем контактной работы в очно-заочной/заочной формах обучения и индивидуальных учебных планах определяется соответствующими учебными планами. Темы, реализуемые в виде контактной работы, определяются преподавателем самостоятельно, исходя из уровня их сложности.

5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8,9	Формы проведения занятий
Введение в NLP, базовая обработка текста и дистанция редактирования	<p>1. Какие основные задачи решает NLP и какие этапы обработки текста вы можете выделить?</p> <p>2. В чем разница между стеммингом и лемматизацией? Какой из этих методов вы бы использовали для русскоязычного текста и почему?</p> <p>3. Что такое расстояние Левенштейна и в каких ситуациях оно полезно?</p> <p>4. Что означает Закон Ципфа в контексте NLP и как он применяется?</p> <p><i>Рекомендуемые источники: п.8, [1]- [3]</i></p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений

Языковые модели и векторная семантика	<p>1. Что такое N-граммы и какая их роль в языковых моделях?</p> <p>2. Какие метрики используются для измерения семантической близости в векторных пространствах?</p> <p>3. Чем отличаются методы Skip-Gram и CBoW в контексте Word2vec?</p> <p>4. Какая роль у WordNet в языковых моделях и в чем его особенности?</p> <p>Рекомендуемые источники: p.8, [1]-[3]</p>	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	<p>1. Что такое эмбединг слова и какова его роль в глубоком обучении?</p> <p>2. Как работает алгоритм Word2vec? В чем разница между Skip-Gram и CBoW?</p> <p>3. Какие есть предобученные модели эмбедингов и для каких задач их можно использовать?</p> <p>4. Какое значение имеет сингулярное разложение при представлении слов в виде векторов?</p> <p>Рекомендуемые источники: p.8, [1]-[3]</p>	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	<p>1. Что такое разметка по частям речи и зачем она нужна в NLP?</p> <p>2. В чем основные преимущества и недостатки скрытых марковских моделей для разметки текста?</p> <p>3. Какие модификации скрытых марковских моделей вы знаете и для каких задач они могут быть полезны?</p> <p>4. Как извлечение именованных сущностей связано с задачей разметки текста?</p> <p>Рекомендуемые источники: p.8, [1]-[3]</p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений
Рекуррентные нейронные сети в NLP	<p>1. В чем основное отличие между RNN, LSTM и GRU?</p> <p>2. Какие задачи в NLP можно решать с помощью Seq2seq моделей на базе RNN?</p> <p>3. Что такое модель ELMo и в чем ее преимущества перед классическими RNN?</p> <p>4. Какие проблемы могут возникнуть при использовании RNN в NLP и как их можно решить?</p> <p>Рекомендуемые источники: p.8, [1]-[3]</p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений

Механизм внимания, BERT	<p>1. Что такое механизм внимания и как он используется в архитектуре Transformer?</p> <p>2. Как работает модель BERT и в чем ее основные преимущества перед другими архитектурами?</p> <p>3. Что такое процедуры pre-training и fine-tuning в контексте работы с моделями на базе BERT?</p> <p>4. Какие существуют предобученные реализации на базе BERT и для каких задач они могут быть использованы?</p> <p>Рекомендуемые источники: п.8, [1]-[3]</p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений
Большие лингвистические модели	<p>1. Что такое Большие лингвистические модели (LLM) и в чем их основное преимущество?</p> <p>2. Какие техники и методы используются для улучшения работы с LLM?</p> <p>3. Что такое промпт-инжиниринг и как он может быть применен в контексте LLM?</p> <p>4. Какие проблемы могут возникнуть при работе с большими лингвистическими моделями и как их можно решить?</p> <p>Рекомендуемые источники: п.8, [1]-[3]</p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений
Технологии работы с LLM	<p>1. Какие методы используются для оценки качества работы больших лингвистических моделей?</p> <p>2. Что такое методы адаптеров и как они применяются в контексте LLM?</p> <p>3. Что такое проблема галлюцинаций у LLM и как архитектура RAG помогает ее решить?</p> <p>4. Какие техники квантизации могут быть применены для улучшения работы LLM?</p> <p>Рекомендуемые источники: п.8, [1]-[3]</p>	Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений

6. Учебно-методическое обеспечение 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	<p>Как влияют стоп-слова на качество анализа текста, и какие методы их идентификации существуют?</p> <p>Какие основные этапы предварительной обработки текста вы бы выделили, и почему каждый из них важен?</p>	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.
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 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

Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания
ПКП-6 Способность разрабатывать, реализовывать и применять методы интеллектуального анализа данных и машинного обучения для автоматизации решения неструктурированных и слабоструктурированных задач экономических предметных областей	Использует знания современных методов интеллектуального анализа данных (в том числе, больших данных) и способы их программной реализации	Знать: подходы к выполнению аналитических исследований в области обработки больших данных на естественном языке. Уметь: выполнять аналитические исследования в области обработки больших данных на естественном языке.	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.
	Осуществляет поиск, сбор, анализ и интерпретацию данных экономических предметных областей с применением методов искусственного интеллекта и машинного обучения	Знать: методы использования текстовых данных для задач в экономических областях с использованием технологий обработки данных на естественном языке.	Используйте методы анализа сентимента для оценки влияния экономических новостей и сообщений на рыночные индексы или активы. Анализируйте эмоциональную окраску новостных статей и их влияние

		Уметь: применять методы использования текстовых данных для задач в экономических областях на основе технологий обработки данных на естественном языке.	на изменение цен и торговые объемы.
	Владеет современными инструментарием искусственного интеллекта и его использованием при разработке и развитии существующих финансово-экономических информационных систем	Знать: ключевые фреймворки и библиотеки для решения задач обработки текста на естественном языке. Уметь: строить модели по обработке текста не естественном языке с использованием современных фреймворков и библиотек и использовать их для развития существующих финансово-экономических информационных систем.	Разработайте прототип чат-бота с функционалом обработки текста на естественном языке для предоставления финансовых консультаций или информации. Используйте современные библиотеки и фреймворки для реализации функционала NLP и проведите тестирование прототипа, оценив его эффективность и точность ответов.

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

- [1]. Иванищева, О. Н. Прикладная лингвистика: учебное пособие / О. Н. Иванищева. — Москва: Русайнс, 2021. — 235 с. — ЭБС BOOK.ru. — URL: <https://book.ru/book/942005> (дата обращения: 15.10.2024). — Текст: электронный.
- [2]. Влавацкая, М. В. Введение в языкознание: учебное пособие / М. В. Влавацкая. — Новосибирск: НГТУ, 2019. — 416 с. — ЭБС Лань. — URL: <https://e.lanbook.com/book/152389>; ЭБС Университетская библиотека online. - URL: <https://biblioclub.ru/index.php?page=book&id=575297> (дата обращения: 15.10.2024). — Текст: электронный.

[3]. Махлина, С. Т. Лингвистика и семиотика: учебник и практикум для вузов / С. Т. Махлина. — Москва: Юрайт, 2024. — 260 с. — (Высшее образование). — ЭБС Юрайт. — URL: <https://urait.ru/bcode/544215> (дата обращения: 15.10.2024). — Текст: электронный.

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 Znanium <http://www.znanium.com>
5. Электронно-библиотечная система издательства «ЮРАЙТ» <https://urait.ru/>
6. Электронно-библиотечная система издательства Проспект <http://ebs.prospekt.org/books>
7. Электронно-библиотечная система издательства Лань <https://e.lanbook.com/>
8. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
9. Электронная библиотека Издательского дома «Гребенников» <https://grebennikon.ru/>
10. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
11. Национальная электронная библиотека <http://нэб.рф/>
12. 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.

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

11.1. Комплект лицензионного программного обеспечения:

1. Пакет офисных программ
2. Антивирус Kaspersky

11.2. 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/>

11.3. Сертифицированные программные и аппаратные средства защиты информации: - не предусмотрены.

11.4. Язык программирования Python 3.8 (или старше)

11.5. Платформа для научных исследований, основанная на языке программирования Python, Anaconda, библиотека PyTorch.

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

Наличие аудитории, оснащенной компьютерной техникой и проектором, с возможностью подключения к сети «Интернет».