

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«Тверской государственный технический университет»
(ТвГТУ)

УТВЕРЖДАЮ

Проректор по учебной работе

Смирнов М.А.
« ____ » _____ 20__ г.

РАБОЧАЯ ПРОГРАММА

Дисциплины, части формируемой участниками образовательных отношений Блока 1 «Дисциплины (модули)» **«Автоматический анализ текстовой информации»**

Направление подготовки бакалавров 09.03.04 Программная инженерия.

Направленность (профиль) — Разработка систем искусственного интеллекта.

Типы задач профессиональной деятельности — производственно-технологический.

Форма обучения – очная.

Факультет информационных технологий.

Кафедра «Программное обеспечение».

Тверь 20

Рабочая программа дисциплины соответствует ОХОП подготовки бакалавров в части требований к результатам обучения по дисциплине и учебному плану.

Разработчик программы

Е.И. Корнеева

Программа рассмотрена и одобрена на заседании кафедры ПО
«_____» 20_____ г., протокол №_____.

Заведующий кафедрой ПО

А. Л. Калабин

Согласовано

Начальник УМО

Е.Э.Наумова

Начальник отдела
комплектования
зональной научной библиотеки

О. Ф. Жмыхова

1. Цель и задачи дисциплины

Целью изучения дисциплины «Автоматический анализ текстовой информации» является формирование у обучающихся представлений о методах и алгоритмах машинного обучения, уяснение основных принципов разработки, внедрения и применения систем машинного обучения для решения прикладных задач.

Задачами дисциплины являются:

- формирование представлений у обучающегося об основных понятиях, используемых в предметной области машинного обучения;
- формирование представлений об областях практического использования методов машинного обучения и проблемах, связанных с их внедрением;
- получение знаний о принципах построения моделей машинного обучения;
- получение практических навыков разработки моделей машинного обучения с использованием библиотек scikit-learn, NumPy, pandas.

2. Место дисциплины в структуре ОП

Дисциплина относится к части, формируемой участниками образовательных отношений, Блока 1 ОП ВО, определяет подготовку бакалавров по направлению Программная инженерия в использовании методов машинного обучения в дальнейшей учебной, научной и профессиональной деятельности. Она требует знания основ математического анализа, линейной алгебры, теории вероятностей и математической статистики.

В результате изучения данной дисциплины студенты будут знать основные методы и алгоритмы машинного обучения, уметь применять их для решения практических задач с помощью средств современных библиотек, с учётом прикладной специфики.

3. Планируемые результаты обучения по дисциплине

3.1. Перечень компетенций, закреплённых за дисциплиной в ОХОП

Компетенция, закрепленная за дисциплиной в ОХОП:

ПК-5. Способен разрабатывать, адаптировать, применять в профессиональной деятельности алгоритмы, программные средства, системы и комплексы обработки данных, методы и алгоритмы машинного обучения, программно-технические платформы, электронные библиотеки, программные оболочки приложений, сетевые технологии для решения задач в сфере искусственного интеллекта и смежных областях.

ПК-6. Способен выбирать, применять и проводить экспериментальную проверку работоспособности программных компонентов систем, включающих модули по созданию искусственного интеллекта.

Индикаторы компетенции, закрепленные за дисциплиной в ОХОП

ИПК-5.5. Умеет структурировать и обрабатывать массивы данных большого объема с помощью специальных автоматизированных инструментов для анализа, прогнозов и принятия решений.

ИПК-5.9. Использует вычислительные методы и алгоритмы для анализа и извлечения значимой информации из больших объемов различного контента с применением различных методов обработки, алгоритмов машинного обучения, интеллектуального анализа данных.

ИПК-6.1. Выбирает, комбинирует и адаптирует существующие программные продукты, для решения необходимых функций, профессиональных задач предприятий или организаций.

ИПК-6.2. Самостоятельно создает на основе стандартных оболочек с привлечением искусственного интеллекта программное обеспечение для решения необходимых функций, профессиональных задач предприятий или организаций.

Показатели оценивания индикаторов достижения компетенций

Знать:

31. основные задачи и уровни автоматической обработки естественного языка (NLP);
32. методы предобработки и векторного представления текстовой информации (Bag of Words, TF-IDF, Word2Vec);
33. алгоритмы классификации, кластеризации и тематического моделирования текстов;
34. метрики оценки качества моделей анализа текста и методы их валидации.

Уметь:

У1. выполнять предварительную обработку текстовых данных (токенизация, лемматизация, удаление стоп-слов) с использованием библиотек Python (NLTK, SpaCy);

У2. строить векторные модели представления текста и применять алгоритмы машинного обучения для решения задач NLP;

Иметь опыт практической подготовки:

ПП1. Интерпретировать результаты анализа и оценивать эффективность разработанных моделей обработки текста.

3.2. Технологии, обеспечивающие формирование компетенций

Проведение лекционных занятий, выполнение лабораторных работ, самостоятельная работа под руководством преподавателя, выполнение курсовой работы.

4. Трудоёмкость дисциплины и виды учебной работы

Таблица 1. Распределение трудоемкости дисциплины по видам учебной работы

Вид учебной работы	Зачётных единиц	Академических часов
Общая трудоемкость дисциплины	3	108
Аудиторные занятия (всего)		52
В том числе:		
Лекции		26
Практические занятия (ПЗ)		не предусмотрены
Лабораторные работы (ЛР)		26
Самостоятельная работа (всего)		56
В том числе:		
Курсовая работа (КР)		30
Курсовой проект (КП)		не предусмотрен
Расчетно-графические работы		не предусмотрены
Реферат		не предусмотрен
Другие виды самостоятельной работы: - подготовка к защите лабораторных работ		26
Контроль текущий и промежуточный (балльно-рейтинговый, экзамен)		0
Контроль текущий и промежуточный (балльно-рейтинговый, зачёт)		0
Практическая подготовка при реализации дисциплины (всего)		0
В том числе:		56
Практические занятия (ПЗ)		не предусмотрены

Лабораторные работы (ЛР)		26
Курсовая работа (КР)		30
Курсовой проект (КП)		не предусмотрен

5. Структура и содержание дисциплины

Структура и содержание дисциплины построены по модульно-блочному принципу. Под модулем дисциплины понимается укрупненная логико-понятийная тема, характеризующаяся общностью использованного понятийно-терминологического аппарата.

5.1. Структура дисциплины

Таблица 2. Модули дисциплины, трудоемкость в часах и виды учебной работы

№	Наименование модуля	Труд-ть (часы)	Лекции	Практич. занятия	Лаб. работы	Сам. работа
1	Основы NLP и предобработка текста	34	8	-	8	18
2	Векторные представления и классификация	34	8	-	8	18
3	Нейросетевые модели и прикладные задачи	40	10	-	8	20
	Всего на дисциплину	108	26	-	26	56

5.2. Содержание дисциплины

МОДУЛЬ 1: «ОСНОВЫ NLP И ПРЕДОБРАБОТКА ТЕКСТА»

Введение в NLP. Задачи автоматической обработки текста (Natural Language Processing). Уровни языка: фонетика, морфология, синтаксис, семантика, прагматика. Основные проблемы обработки естественного языка: неоднозначность, неологизмы, сленг.

Предобработка текстовых данных. Стандартный пайплайн предобработки (NLP pipeline). Токенизация: разбиение на слова и предложения. Удаление стоп-слов и пунктуации. Нормализация текста:

стемминг (алгоритм Портера) и лемматизация. Очистка от шума (HTML-теги, спецсимволы). Использование регулярных выражений (RegEx) для поиска и обработки текстовых паттернов. Обзор популярных библиотек: NLTK, SpaCy, PyMorphy2.

МОДУЛЬ 2: «ВЕКТОРНЫЕ ПРЕДСТАВЛЕНИЯ И КЛАССИФИКАЦИЯ»

Векторизация текста. Проблема представления текста в машинном виде. Модель «Мешок слов» (Bag of Words). Понятие N-грамм. Матрица «термин-документ». Взвешивание терминов: TF-IDF (Term Frequency - Inverse Document Frequency). Разреженные матрицы и работа с ними.

Дистрибутивная семантика. Векторные представления слов (Word Embeddings). Гипотеза дистрибутивной семантики. Архитектуры Word2Vec (CBOW, Skip-gram). Оценка качества эмбеддингов. Арифметика векторов. Модели GloVe и FastText.

Классификация текстов. Постановка задачи классификации (бинарная, многоклассовая). Применение алгоритмов машинного обучения: Наивный байесовский классификатор (Naive Bayes), Логистическая регрессия, Метод опорных векторов (SVM). Метрики оценки качества: Accuracy, Precision, Recall, F1-score, ROC-AUC.

МОДУЛЬ 3: «НЕЙРОСЕТЕВЫЕ МОДЕЛИ И ПРИКЛАДНЫЕ ЗАДАЧИ»

Нейронные сети в NLP. Ограничения классических моделей. Рекуррентные нейронные сети (RNN) для обработки последовательностей. Проблема затухающего градиента. Архитектуры LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit).

Современные архитектуры. Механизм внимания (Attention). Понятие Трансформеров (Transformers). Обзор архитектур BERT (Bidirectional Encoder Representations from Transformers) и GPT (Generative Pre-trained Transformer). Использование предобученных моделей (Transfer Learning).

Прикладные задачи. Анализ тональности (Sentiment Analysis). Тематическое моделирование (Topic Modeling): LDA, BigARTM. Извлечение именованных сущностей (NER). Машинный перевод и генерация текста..

Специальные темы. Работа с временными рядами. Обнаружение аномалий (anomaly detection). Системы рекомендаций: коллаборативная фильтрация, content-based методы. Transfer learning и domain adaptation. Интерпретируемость моделей машинного обучения. Этические аспекты и bias в машинном обучении. MLOps: жизненный цикл моделей машинного обучения в production.

5.3. Лабораторные работы

Таблица 3. Лабораторные работы и их трудоемкость

Модули. Цели лабораторных работ	Примерная тематика лабораторных работ	Трудоемкость в часах
Модуль 1 Цель: изучение методов сбора и предварительной обработки текстовых данных.	Использование регулярных выражений для очистки и нормализации текста.	4
	Токенизация и лемматизация текстов с использованием библиотек NLTK/SpaCy. Частотный анализ.	4
Модуль 2 Цель: освоение алгоритмов векторизации и классификации текстов.	Построение признакового пространства методами Bag of Words и TF-IDF	4
	Обучение классификаторов (Naive Bayes, SVM) для рубрикации текстов или определения спама.	4
Модуль 3 Цель: освоение ансамблевых методов, методов кластеризации и снижения размерности	Обучение и визуализация векторных представлений слов (Word2Vec).	5
	Разработка модели анализа тональности (Sentiment Analysis) на основе рекуррентных сетей или трансформеров.	5

5.4. Практические занятия

Учебным планом практические занятия не предусмотрены.

6. Самостоятельная работа обучающихся и текущий контроль успеваемости

6.1. Цели самостоятельной работы

Формирование способностей к самостоятельному познанию и обучению, поиску литературы, обобщению, оформлению и представлению полученных результатов, их критическому анализу, поиску новых и неординарных решений, аргументированному отстаиванию своих предложений, умений подготовки выступлений и ведения дискуссий.

6.2. Организация и содержание самостоятельной работы

Самостоятельная работа заключается в изучении отдельных тем курса по заданию преподавателя по рекомендуемой им учебной литературе, в подготовке к лабораторным работам, к текущему контролю успеваемости, экзамену, зачёту с оценкой, в выполнении курсовой работы.

В рамках дисциплины выполняется **15 лабораторных работ** (5-й семестр — 5 работ, 6-й семестр — 10 работ), которые защищаются посредством тестирования или устным опросом (по желанию обучающегося). Максимальная оценка за каждую выполненную лабораторную работу – 5 баллов, минимальная – 3 балла.

Выполнение всех лабораторных работ обязательно. В случае невыполнения лабораторной работы по уважительной причине студент имеет право выполнить письменный реферат, по согласованной с преподавателем теме по модулю, по которому пропущена лабораторная работа. Возможная тематическая направленность реферативной работы для каждого учебно-образовательного модуля представлена в следующей таблице:

7. Учебно-методическое и информационное обеспечение дисциплины

7.1. Основная литература по дисциплине

1. Сотник, С.Л. Проектирование систем искусственного интеллекта : учебное пособие / С.Л. Сотник. - 3-е изд. - Москва : Интернет-Университет Информационных Технологий (ИНТУИТ) : Ай Пи Ар Медиа, 2021. - Текст : электронный. - Режим доступа: по подписке. - Дата обращения: 07.07.2022. - ЦОР IPR SMART. - ISBN 978-5-4497-0868-7. - URL: <https://www.iprbookshop.ru/102054.html> . - (ID=145140-0)
2. Нугуманова, А. Б. Автоматизированная обработка текстовых массивов : учебник и практикум для вузов / А. Б. Нугуманова. — 2-е изд. — Москва : Издательство Юрайт, 2024. — 82 с. — (Высшее образование). — ISBN 978-5-534-20738-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/558668> (дата обращения: 22.12.2025). - (ID=189533-0)

3. Борисова, И. В. Цифровые методы обработки информации : учебное пособие / И. В. Борисова. — Новосибирск : Новосибирский государственный технический университет, 2014. — 139 с. — ISBN 978-5-7782-2448-3. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/45061.html> (дата обращения: 22.12.2025). — Режим доступа: для авторизир. пользователей. - (ID=189534-0)
4. Гутгарц, Р. Д. Проектирование автоматизированных систем обработки информации и управления : учебник для вузов / Р. Д. Гутгарц. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 351 с. — (Высшее образование). — ISBN 978-5-534-15761-1. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/565010> (дата обращения: 22.12.2025). - (ID=143955-0)

7.2. Дополнительная литература по дисциплине

1. Качановский, Ю. П. Технологии обработки информации в текстовом процессоре Microsoft Word : методические указания к проведению лабораторной работы по курсу «Информатика» / Ю. П. Качановский, А. С. Широков. — Липецк : Липецкий государственный технический университет, ЭБС АСВ, 2014. — 35 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/55164.html> (дата обращения: 22.12.2025). — Режим доступа: для авторизир. пользователей. - (ID=189537-0)
2. Толмачев, С.Г. Основы искусственного интеллекта : учебное пособие для вузов / С.Г. Толмачев; Балтийский государственный технический университет «Военмех» имени Д.Ф. Устинова. - Санкт-Петербург : Балтийский государственный технический университет «Военмех» имени Д.Ф. Устинова, 2017. - Текст : электронный. - Режим доступа: по подписке. - Дата обращения: 07.07.2022. - ЭБС Лань. - ISBN 978-5-906920-53-9. - URL: <https://e.lanbook.com/book/121872> . - (ID=147570-0)
3. Масленникова, О.Е. Основы искусственного интеллекта : учебное пособие для вузов по специальности 050202.65 - "Информатика" / О.Е. Масленникова, И.В. Гаврилова. - 2-е изд., стер. - Москва : Флинта, 2013. - Текст : электронный. - Дата обращения: 07.07.2022. - Режим доступа: по подписке. - ЭБС Университетская библиотека онлайн. - ISBN 978-5-9765-1602-1. - URL:

https://biblioclub.ru/index.php?page=book_view_red&book_id=363418. - (ID=113593-0)

4. Галыгина, И.В. Основы искусственного интеллекта. Лабораторный практикум : учебное пособие для вузов / И.В. Галыгина, Л.В. Галыгина. - 2-е изд., стер. - Санкт-Петербург [и др.] : Лань, 2024. - Текст : электронный. - Режим доступа: по подписке. - Дата обращения: 20.09.2022. - ЭБС Лань. - ISBN 978-5-507-48767-7. - URL: <https://e.lanbook.com/book/362927> . - (ID=150122-0)
5. Баранникова, И. В. Теоретические основы автоматизированной обработки информации и управления. Специальные функции MS Excel : лабораторный практикум / И. В. Баранникова, Е. С. Могирева, О. Г. Харахан. — Москва : Издательский Дом МИСиС, 2018. — 61 с. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/78543.html> (дата обращения: 22.12.2025). — Режим доступа: для авторизир. пользователей. - (ID=189536-0)
6. Батура, Т. В. Математическая лингвистика и автоматическая обработка текстов на естественном языке : учебное пособие / Т. В. Батура. — Новосибирск : Новосибирский государственный университет, 2016. — 166 с. — ISBN 978-5-4437-0548-4. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/93489.html> (дата обращения: 22.12.2025). — Режим доступа: для авторизир. пользователей. - (ID=189535-0)

7.3. Методические материалы

1. Вопросы к зачету по дисциплинам "Алгоритмы и структуры данных" направления подготовки 09.03.04 Программная инженерия. Профиль: Разработка программно-информационных систем : в составе учебно-методического комплекса / Каф. Программное обеспечение вычислительной техники ; разраб. А.А. Мальков. - Тверь : ТвГТУ, 2017. - (УМК-В). - Текст : электронный. - Сервер. - (ID=131108-0)
2. Оценочные средства промежуточной аттестации в форме курсовой работы по дисциплине "Теория алгоритмов" направления подготовки 09.03.04 Программная инженерия. Профиль: Разработка программно-информационных систем : в составе учебно-методического комплекса / Каф. Программное обеспечение вычислительной техники ; разраб. Б.Н. Карлов. - Тверь : ТвГТУ,

2017. - (УМК-КП). - Текст : электронный. - Сервер. - URL: <https://elib.tstu.tver.ru/MegaPro/GetDoc/Megapro/131578> . - (ID=131578-0)
3. Оценочные средства промежуточной аттестации в форме зачета по дисциплине "Теория алгоритмов" направления подготовки 09.03.04 Программная инженерия. Профиль: Разработка программно-информационных систем : в составе учебно-методического комплекса / Каф. Программное обеспечение вычислительной техники ; разраб. Б.Н. Карлов. - Тверь : ТвГТУ, 2017. - (УМК-В). - Текст : электронный. - Сервер. - URL: <https://elib.tstu.tver.ru/MegaPro/GetDoc/Megapro/131577> . - (ID=131577-0)
4. Вопросы по дисциплине "Теория алгоритмов" направления подготовки 09.03.04 Программная инженерия. Профиль: Разработка программно-информационных систем : в составе учебно-методического комплекса / Каф. Программное обеспечение вычислительной техники ; разраб. Б.Н. Карлов. - Тверь : ТвГТУ, 2017. - (УМК-В). - Текст : электронный. - Сервер. - URL: <https://elib.tstu.tver.ru/MegaPro/GetDoc/Megapro/131579> . - (ID=131579-0)
5. Учебно-методический комплекс дисциплины обязательной части Блока 1 "Дисциплины (модули)" "Автоматический анализ текстовой информации". Направление подготовки бакалавров 09.03.04 Программная инженерия, Направленность (профиль) – Разработка систем искусственного интеллекта : ФГОС 3++ / сост. Е.И. Корнеева ; Каф. Программное обеспечение. - Тверь, 2025. - (УМК). - Текст : электронный. - URL: <https://elib.tstu.tver.ru/MegaPro/GetDoc/Megapro/189531> . - (ID=189531-0)

7.4. Программное обеспечение по дисциплине

Операционная система Microsoft Windows: лицензии № ICM-176609 и № ICM-176613 (Azure Dev Tools for Teaching).

Microsoft Office 2007 Russian Academic: OPEN No Level: лицензия № 41902814.

Python 3.8+ (свободная лицензия PSF).

Среды разработки:

- Jupyter Notebook / JupyterLab (BSD лицензия)
- PyCharm Community Edition (Apache 2.0)
- VS Code (MIT лицензия)

7.5. Специализированные базы данных, справочные системы, электронно-библиотечные системы, профессиональные порталы в Интернет

ЭБС и лицензионные ресурсы ТвГТУ размещены:

1. Ресурсы: <https://lib.tstu.tver.ru/header/obr-res>
2. ЭБ ТвГТУ: <https://elib.tstu.tver.ru/MegaPro/Web>
3. ЭБС "Лань": <https://e.lanbook.com/>
4. ЭБС "Университетская библиотека онлайн": <https://biblioclub.ru/>
5. Национальная электронная библиотека: <https://rusneb.ru>
6. ЦОР IPRSmart: <https://www.iprbookshop.ru/>
7. Электронная образовательная платформа "Юрайт": <https://urait.ru/>
8. Научная электронная библиотека eLIBRARY: <https://elibrary.ru/>
9. Информационная система "ТЕХНОМАТИВ". Конфигурация "МАКСИМУМ": сетевая версия (годовое обновление) : [нормативно-технические, нормативно-правовые и руководящие документы (ГОСТы, РД, СНИПы и др.]. Диск 1, 2, 3, 4. - М. :Технорматив, 2014. - (Документация для профессионалов). - CD. - Текст : электронный. - 119600 р. – (105501-1)
10. База данных учебно-методических комплексов: <https://lib.tstu.tver.ru/header/umk.html>

Профессиональные ресурсы:

11. Kaggle: <https://www.kaggle.com/>
12. Papers with Code: <https://paperswithcode.com/>
13. Scikit-learn Documentation: <https://scikit-learn.org/>
14. UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/>

УМК размещен:<https://elib.tstu.tver.ru/MegaPro/GetDoc/Megapro/189531>

8. Материально-техническое обеспечение

При изучении дисциплины «Автоматический анализ текстовой информации» используются современные средства обучения: наглядные пособия, диаграммы, схемы.

Возможна демонстрация лекционного материала с помощью оверхед-проектора (кодоскопа) и мультипроектора.

Вуз имеет лабораторию для реализации лабораторного практикума по дисциплине «Автоматический анализ текстовой информации», учебный класс для проведения самостоятельной работы по курсу, оснащённый современной компьютерной и офисной техникой, необходимым программным обеспечением, электронными учебными пособиями, имеющий безлимитный выход в глобальную сеть; аудиторию для

проведения семинарских занятий, практикумов и презентаций студенческих работ, оснащённую аудиовизуальной техникой.

9. Оценочные средства для проведения промежуточной аттестации

9.1. Оценочные средства для проведения промежуточной аттестации в форме экзамена

Учебным планом не предусмотрен.

9.2. Оценочные средства для проведения промежуточной аттестации в форме зачета с оценкой

1. Шкала оценивания промежуточной аттестации в форме зачёта – «зачтено», «не зачтено».
2. Вид промежуточной аттестации в форме зачёта.
Вид промежуточной аттестации устанавливается преподавателем: по результатам текущего контроля знаний обучающегося без дополнительных контрольных испытаний или с выполнением дополнительного итогового контрольного испытания при наличии у студентов задолженностей в текущем контроле.
3. Для дополнительного итогового контрольного испытания студенту в обязательном порядке предоставляется: база заданий, предназначенных для предъявления обучающемуся на дополнительном итоговом контрольном испытании (типовы образец задания приводится в Приложении), задание выполняется письменно.
4. База заданий, предъявляемая обучающимся на дополнительном итоговом контрольном испытании:

1. В чем заключается задача автоматической обработки текста (NLP) и какие уровни языка она затрагивает?
2. Опишите стандартный пайплайн предобработки текста: токенизация, лемматизация, удаление стоп-слов.
3. Что такое регулярные выражения и как они применяются при очистке текстовых данных?
4. В чем различие между стеммингом и лемматизацией? Приведите примеры.
5. Опишите модель "Мешок слов" (Bag of Words). В чем её преимущества и недостатки?
6. Объясните принцип работы метрики TF-IDF. Что означает каждое из слагаемых?
7. Что такое N-граммы и как выбор N влияет на размерность признакового пространства?

8. В чём заключается гипотеза дистрибутивной семантики?
9. Опишите архитектуру Word2Vec (CBOW и Skip-gram).
10. Как оценивается качество векторных представлений слов (Word Embeddings)?
11. Опишите применение Наивного байесовского классификатора для задачи классификации текстов.
12. Какие метрики используются для оценки качества классификации текстов (Precision, Recall, F1-score)?
13. В чём особенность применения рекуррентных нейронных сетей (RNN) для анализа текста?
14. Что такое механизм внимания (Attention) и какую проблему он решает?
15. Опишите основные принципы работы моделей на базе Трансформеров (BERT, GPT).
16. Что такое тематическое моделирование? Опишите суть алгоритма LDA.
17. Как решается задача анализа тональности (Sentiment Analysis)?
18. Что такое именованные сущности (NER) и как происходит их извлечение?
19. Какие существуют методы борьбы с дисбалансом классов в текстовых датасетах?
20. Приведите примеры современных прикладных задач NLP в бизнесе.

Методические материалы определяют процедуру проведения дополнительного итогового испытания и проставления зачёта.

Критерии выполнения контрольного испытания и условия проставления зачёта: критерии оценки для категории «знать» (количественный критерий):

Ниже базового – 0 баллов,

Базовый уровень – 1 балл.

Критерии оценки для категории «уметь» (бинарный критерий):

Отсутствие умения – 0 баллов,

Наличие умения – 1 балл.

Критерии итоговой оценки за зачёт:

«зачтено» – при сумме баллов 2 или 3;

«не зачтено» – при сумме баллов 0 или 1.

Число заданий для дополнительного итогового контрольного испытания –

20. Число вопросов – 3. Продолжительность – 60 минут.

4. При промежуточной аттестации без выполнения дополнительного итогового контрольного испытания студенту в обязательном порядке описываются критерии проставления зачёта: «зачтено» –

выставляется обучающемуся при условии выполнения им всех контрольных мероприятий.

9.3. Оценочные средства для проведения промежуточной аттестации в форме курсовой работы

1. Шкала оценивания курсовой работы — «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

2. Примерная тематика курсовой работы.

Каждому обучающемуся выдаётся индивидуальный вариант для разработки программного обеспечения. Студент по согласованию с преподавателем может самостоятельно выбрать тему для курсовой работы.

3. Критерии оценки качества выполнения как по отдельным разделам курсовой работы, так и работы в целом.

Разделы курсовой работы по дисциплине «Методы и алгоритмы машинного обучения»:

№ раздела	Наименование раздела	Баллы по шкале уровня
-	Нормативные ссылки	Выше базового — 2 Базовый — 1 Ниже базового — 0
-	Термины и определения	Выше базового — 2 Базовый — 1 Ниже базового — 0
-	Введение	Выше базового — 2 Базовый — 1 Ниже базового — 0
1	Общая часть (обзор литературы по теме курсовой работы, описание данных, постановка задачи)	Выше базового — 6 Базовый — 3 Ниже базового — 0
2	Специальная часть (описание алгоритмов, разработка и обучение моделей, эксперименты)	Выше базового — 6 Базовый — 3 Ниже базового — 0
-	Заключение	Выше базового — 2 Базовый — 1 Ниже базового — 0
-	Список использованных источников	Выше базового — 2 Базовый — 1 Ниже базового — 0
-	Приложения (исходный код, дополнительные графики и таблицы)	Выше базового — 6 Базовый — 3 Ниже базового — 0

Критерии итоговой оценки за курсовую работу:

- «отлично» – при сумме баллов от 24 до 28;
- «хорошо» – при сумме баллов от 19 до 23;
- «удовлетворительно» – при сумме баллов от 14 до 18;
- «неудовлетворительно» – при сумме баллов менее 14.

Методические материалы, определяющие процедуру выполнения и представления работы и технологию её оценивания.

Требования и методические указания по структуре, содержанию и выполнению работы, а также критерии оценки оформлены в качестве отдельно выпущенного документа.

Курсовая работа состоит из титульного листа, содержания, нормативных ссылок, терминов и определений, сокращений, введения, основной части, экспериментальной части, заключения, списка использованных источников и приложений. Текст должен быть структурирован, содержать рисунки и таблицы. Рисунки и таблицы должны располагаться сразу после ссылки на них в тексте таким образом, чтобы их можно было рассматривать без поворота курсовой работы. Если это сложно, то допускается поворот по часовой стрелке.

Если таблицу приходится переносить на следующую страницу, то помещают слова: «продолжение табл.» с указанием номера справа, графы таблицы пронумеровывают и повторяют их нумерацию на следующей странице. Заголовок таблицы не повторяют.

В введении необходимо отразить актуальность темы исследования, цель и задачи курсовой работы. Объем должен составлять 1-2 страницы.

Теоретическая часть должна содержать обзор актуальных литературных и нормативных источников выбранного объекта курсовой работы.

В практической части необходимо отразить:

- проектирование структуры ПО;
- проектирование интерфейса ПО;
- ход реализации разрабатываемого ПО;
- примеры тестирования ПО;

В заключении необходимо раскрыть особенности отображения в курсовой работе поставленных задач. Объем должен составлять 1-2 страницы.

Список использованных источников должен содержать не менее 10 наименований (книг, журналов, электронных ресурсов и др.).

Дополнительные процедурные сведения:

а) Студенты выбирают тему для курсовой работы самостоятельно из предложенного списка и согласовывают свой выбор с преподавателем в течение первых двух недель обучения. К середине семестра на проверку

представляется общая часть курсовой работы, за две недели до защиты – окончательный вариант.

б) проверку и оценку работы осуществляет руководитель, который доводит до сведения обучающегося достоинства и недостатки курсовой работы, и ее оценку. Оценка проставляется в зачётную книжку обучающегося и ведомость для курсовой работы. Если обучающийся не согласен с оценкой руководителя, проводится защита работы перед комиссией, которую назначает заведующий кафедрой;

в) защита курсовой работы проводится в течение двух последних недель семестра и выполняется в форме устной защиты в виде доклада на 5-7 минут с последующим ответом на поставленные вопросы, в ходе которых выясняется глубина знаний студента и самостоятельность выполнения работы;

г) работа не подлежит обязательному внешнему рецензированию.

В процессе выполнения обучающимся курсовой работы руководитель осуществляет систематическое консультирование.

Оптимальный объем курсовой работы 20-30 страниц машинописного текста (не включая приложения), набранного 12-14 шрифтом через 1.5 интервала на листах формата А4 с одной стороны. Поля должны составлять 20 мм сверху и снизу, 30 мм слева и 15 мм справа.

Источники использованной литературы должны оформляться согласно ГОСТ 7.1-2003 «Библиографическая запись. Библиографическое описание. Общие требования и правила составления». Список источников следует составлять в порядке упоминания их в тексте. Ссылки на источники должны приводиться по тексту в квадратных скобках.

Нумерация страниц курсовой работы должна быть сквозной. Первой страницей является титульный лист, на нем номер страницы не ставится, второй - содержание и т.д. Номер страницы проставляется арабскими цифрами снизу страницы, посередине. Приложения необходимо включать в сквозную нумерацию.

Защита курсовой работы проводится в течение двух последних недель семестра и выполняется в форме устной защиты в виде доклада на 5-7 минут с последующим ответом на поставленные вопросы, в ходе которых выясняется глубина знаний студента и самостоятельность выполнения работы;

Работа не подлежит обязательному внешнему рецензированию.

В процессе выполнения обучающимся курсовой работы руководитель осуществляет систематическое консультирование.

9.4. Оценочные средства для проведения промежуточной аттестации в форме курсового проекта

Учебным планом курсовой проект по дисциплине не предусмотрен.

10. Методические рекомендации по организации изучения дисциплины

Студенты перед началом изучения дисциплины ознакомлены с системами кредитных единиц и балльно-рейтинговой оценки.

Студенты, изучающие дисциплину, обеспечиваются электронными изданиями или доступом к ним, учебно-методическим комплексом по дисциплине, включая методические указания к выполнению лабораторных, курсовых работ и всех видов самостоятельной работы.

В учебный процесс рекомендуется внедрение субъект-субъектной педагогической технологии, при которой в расписании каждого преподавателя определяется время консультаций студентов по закрепленному за ним модулю дисциплины.

11. Внесение изменений и дополнений в рабочую программу дисциплины

Содержание рабочих программ дисциплин ежегодно обновляется протоколами заседаний кафедры по утвержденной «Положением о структуре, содержании и оформлении рабочих программ дисциплин по образовательным программам, соответствующим ФГОС ВО с учетом профессиональных стандартов» форме.

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тверской государственный технический университет»

Направление подготовки бакалавров 09.03.04 Программная инженерия
Профиль – Разработка систем искусственного интеллекта.
Кафедра «Программное обеспечение»
Дисциплина «Автоматический анализ текстовой информации»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Вопрос для проверки уровня «ЗНАТЬ» – 0 или 1 или 2 балла:
Линейная регрессия: метод наименьших квадратов, градиентный спуск.

2. Задание для проверки уровня «УМЕТЬ» - 0 или 1 балл:
Дана матрица ошибок бинарного классификатора. Вычислите accuracy, precision, recall и F1-score. Интерпретируйте результаты.

3. Задание для проверки уровня «УМЕТЬ» – 0 или 2 балла:
Реализуйте алгоритм нормализации признаков (min-max scaling).
Объясните, когда следует применять нормализацию, а когда — стандартизацию.

Критерии итоговой оценки за экзамен:

«отлично» - при сумме баллов 4 или 5;
«хорошо» - при сумме баллов 3;
«удовлетворительно» - при сумме баллов 2;
«неудовлетворительно» - при сумме баллов 0 или 1.

Составитель: _____ Е.И. Корнеева

Заведующий кафедрой: д.ф.-м.н. _____ А.Л.Калабин