Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Филологический факультет

УТВЕРЖДЕНО:
Декан
И.В. Тубалова

Оценочные материалы по дисциплине

Машинное обучение

по направлению подготовки

45.04.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
Компьютерная и когнитивная лингвистика

Форма обучения
Очная

Квалификация
Магистр

Год приема
2025

СОГЛАСОВАНО:
Руководитель ОП
З.И. Резанова

Председатель УМК
Ю.А. Тихомирова

Томск – 2025

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен решать профессиональные задачи, применяя основные понятия, категории и положения лингвистических теорий и актуальные концепции в области лингвистики.

ПК-1 Способен проводить самостоятельные исследования и получать новые научные результаты в области междисциплинарных лингвистических исследований.

ПК-3 Способен разрабатывать системы автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистические компоненты электронных ресурсов и интеллектуальных электронных систем (лингвистические корпуса, словари, онтологии, базы данных).

ПК-4 Способен разрабатывать проекты прикладной направленности в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.2 Решает профессиональные задачи, применяя основные понятия, категории и положения лингвистических теорий

ИПК-1.3 Последовательно реализует исследовательскую программу, получает новые научные результаты

ИПК-3.1 Разрабатывает системы автоматической обработки звучащей речи и письменного текста на естественном языке

ИПК-3.2 Разрабатывает лингвистические компоненты электронных ресурсов (лингвистические корпуса, словари)

ИПК-4.1 Формулирует цель проекта прикладной направленности в области когнитивной и компьютерной лингвистики, обосновывает необходимость применения современных технических средств и информационных технологий, в том числе в области искусственного интеллекта

ИПК-4.3 Обеспечивает выполнение проекта в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта, в соответствии с установленными целями, сроками и затратами

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

— домашние работы.

Примеры заданий домашней работы:

1. (ИОПК-1.2, ИПК-1.3)Дан датасет Pima Indians Diabetes Database (National Institute of Diabetes and Digestive and Kidney Diseases), в котором собраны медицинские данные 768 женщин племени индейцев Пима — их показатели глюкозы, ИМТ, вес, возраст и т.д, — и данные о наличии у них диабета. Сформулируйте задачу, которую вы будете решать с использованием машинного обучения. Изучите датасет, проанализируйте данные в нем при помощи имеющихся у вас навыков Python. Обучите изученные вами модели машинного обучения на данном датасете; полученные обученные модели оцените при помощи матрицы ошибок (sklearn.metrics.confusion_matrix). Сравните модели, прокомментируйте полученные результаты.
2. (ИПК-3.1, ИПК-3.2, ИПК-4.1, ИПК-4.3) Создайте и обучите не менее 4 изученных на занятиях классификатора текстов из созданного на занятии датасета. Сформулируйте лингвистическую задачу, которую вы будете решать с использованием машинного обучения; обоснуйте необходимость его применения.

Используйте объект GridsearchCV для создания сетки гиперпараметров и выбора лучших из них. Выведите лучшие гиперпараметры для каждого классификатора в отдельной ячейке. Проведите оценку классификаторов на этих гиперпараметрах, сравните их, проанализируйте результаты (используйте метрики accuracy, precision, recall, F-score).

Критерии оценивания:

Оценка	Критерии
Отлично (90–100 баллов)	Задание выполнено без отклонений от требований и без недочетов или их мало и они незначительны
Хорошо (70–89 баллов)	Задание выполнено без отклонений от требований и с приемлемым количеством незначительных несистемных недочетов
Удовлетворительно (50–69 баллов)	Задание выполнено с несколькими значительными недочетами (применительно к требованиям) или незначительных недочетов много и/или они системны
Неудовлетворительно (менее 50 баллов)	Задание не выполнено либо не выполнены требования к нему или недочетов больше, чем правильного текста

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Зачет с оценкой принимается в форме проекта.

Для зачета необходимо выполнить следующее задание:

Реализовать программный код средней сложности на языке Python по текстовому описанию, которое предполагает формулировку лингвистической задачи, анализ лингвистических данных, обоснование применения машинного обучения для решения данной лингвистической задачи, разработку системы машинного обучения, обучение модели машинного обучения и анализ результатов работы этой модели в контексте поставленной цели. Текстовое описание задания имеет следующее содержание:

«Примените один из рассмотренных в ходе курса методов машинного обучения для решения задачи классификации на текстовых данных, которые не рассматривались в ходе курса (собранные самостоятельно в ходе курса Язык программирования Python или найденные в интернете).

Опишите задачу и объясните выбор метода решения задачи в комментариях к коду. Оцените получившийся классификатор с помощью метрик, изученных в ходе курса.

Прикрепите архив с данными или ссылку на них.»

Проект сдается в письменном виде в соответствующем элементе системы LMS iDo. При сдаче проекта оценивается выполняемость программного кода, умение решать профессиональные задачи, применяя основные понятия, категории и положения лингвистических теорий (ИОПК-1.2), умение последовательно реализовать исследовательскую программу, получать новые научные результаты (ИПК-1.3), разрабатывать системы машинного обучения для автоматической обработки звучащей речи и письменного текста на естественном языке (ИПК-3.1), которые могут являться компонентами электронных ресурсов (ИПК-3.2), в рамках проектной деятельности в

соответствии с установленными целями, сроками и затратами (ИПК-4.3). Оценивается умение формулировать лингвистическую задачу, для решения которой используется машинное обучение — цель проекта, а также умение обосновывать необходимость применения соответствующих технических средств и информационных технологий (ИПК-4.1).

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» ставится при условиях: классификатор реализован без ошибок или с минимальными недочетами, не влияющими на результат, решает задачу с высокой точностью, код структурирован, содержит подробные комментарии с четким описанием задачи и обоснованием метода; метрики корректно применены и интерпретированы; своевременное выполнение домашних заданий и посещаемость не менее 80% занятий;

Оценка «хорошо» ставится при условиях: классификатор работает с незначительными недочетами, код читаем, комментарии объясняют задачу и выбор метода, но могут быть недостаточно подробными; метрики применены, но их интерпретация недостаточна; выполнение домашних заданий, посещаемость не менее 60% занятий.

Оценка «удовлетворительно» ставится при условиях: классификатор решает задачу с заметными недочетами, код минимально структурирован, комментарии недостаточны; метрики применены с ошибками или без интерпретации; выполнение домашних заданий, посещаемость менее 60% занятий.

Оценка «неудовлетворительно» ставится при условиях: проект не сдан, классификатор не работает или не решает задачу, код отсутствует, неструктурирован и не содержит комментариев; метрики не применены и не интерпретированы; домашние задания не выполнены.

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

1. (ИОПК-1.2) Сопоставьте понятие в лингвистике с профессиональной задачей в области компьютерной лингвистики, решаемой с применением машинного обучения:

a. генеративная грамматика
b. корпусная лингвистика
c. семантический анализ

1. построение модели для анализа синтаксической структуры предложений
2. создание датасета для обучения модели на распознавание именованных сущностей
3. моделирование смысла предложений для задач вопросно-ответных систем

2. (ИПК-1.3) В какой последовательности осуществляется реализация исследовательской деятельности при разработке системы машинного обучения?

a. обучение модели
b. настройка гиперпараметров модели
c. разработка датасета
d. сбор данных
e. формулировка лингвистической задачи
f. оценка моделей при помощи метрик
g. анализ полученных результатов
h. выбор модели

3. (ИПК-3.1) Какие библиотеки языка программирования Python могут быть использованы при разработке системы автоматической обработки звучащей речи и письменного текста на естественном языке неспосредственно для произведения предобработки текста?

a. scikit-learn
b. pandas
c. nltk
d. datetime
e. natasha

4. (ИПК-3.2) Какие методы кодирования текстовых данных могут быть использованы при обучении моделей в рамках разработки систем машинного обучения как лингвистических компоненты электронных ресурсов?

a. One-hot encoding
b. Bag-of-words
c. word-2-vec
d. gpt-3

5. (ИПК-4.1) Сопоставьте набор признаков в датасете с лингвистической задачей, решаемой при помощи машинного обучения:

a. текст новости; заголовок новости
b. текст отзыва пользователя на фильм; положительная, отрицательная или нейтральная оценка фильма
c. текст новости, тема новости

1. классификация текстов
2. семантический анализ
3. генерация текста на основе текста

6. (ИПК-4.3) Сопоставьте метод машинного обучения с лингвистической задачей, решаемой при помощи машинного обучения, опираясь на обоснованность его применения:

a. классификация текстов
b. кластеризация текстов
c. генерация текста на основе текста

1. метод ближайших соседей
2. дерево решений
3. нейронные сети

Ключи:

1. a-1; b-2; c-3 2. e,c,d,h,a,b,f,g; 3. a,b,c,e 4. a,b,c 5. a-3; b-2; c-1 6. a-2; b-1; c-3

Информация о разработчиках

Аишева Динара Арамановна, ассистент каф. общей, компьютерной и когнитивной лингвистики;

Шамигов Федор Федорович, специалист по учебно-методической работе.