МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ЛУГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ ВЛАДИМИРА ДАЛЯ»

Краснодонский факультет инженерии и менеджмента (филиал)
Кафедра информационных технологий и транспорта

УТВЕРЖДАЮ:
Директор
_____ Панайотов К.К.
(подпись)
«14» марта 2025 года

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
по учебной дисциплине
Технологии работы с естественным языком
(наименование учебной дисциплины, практики)

09.04.01 Информатика и вычислительная техника
(код и наименование направления подготовки (специальности))

«Интеллектуальные системы
в производственно-транспортных комплексах»
наименование профиля подготовки (специальности, магистерской программы); при отсутствии ставится прочерк)

Разработчик(разработчики):
доцент _____ Бихдрикер А. С.
(подпись)

ФОС рассмотрен и одобрен на заседании кафедры информационных технологий и транспорта от «26» февраля 2025 г., протокол № 7

Заведующий кафедрой
информационных
технологий и транспорта _____ Верительник Е. А.
(подпись)

Краснодон 2025

Комплект оценочных материалов по дисциплине
«Технологии работы с естественным языком»

Задания закрытого типа

Задания закрытого типа на выбор правильного ответа

1. *Выберите один правильный ответ.*
Какая задача NLP связана с определением части речи каждого слова в тексте?
А) Токенизация.
Б) Стемминг.
В) Лемматизация.
Г) Part-of-Speech Tagging (POS-Tagging).
Правильный ответ: Г
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. *Выберите один правильный ответ.*
Что такое токенизация в NLP?
А) Преобразование текста в нижний регистр.
Б) Удаление стоп-слов из текста.
В) Разделение текста на отдельные слова или фразы.
Г) Приведение слов к их словарной форме.
Правильный ответ: В
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. *Выберите один правильный ответ.*
Что такое стемминг (stemming)?
А) Определение эмоций, выраженных в тексте.
Б) Приведение слов к их корневой форме путем отбрасывания суффиксов и приставок.
В) Разделение текста на предложения.
Г) Замена слов синонимами.
Правильный ответ: Б
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4. *Выберите один правильный ответ.*
Что такое “Word Embedding”?
А) Метод сжатия текста;
Б) Представление слов в виде векторов в многомерном пространстве, отражающее их семантические отношения.
В) Метод автоматического перевода текста с одного языка на другой.
Г) Алгоритм для исправления грамматических ошибок.
Правильный ответ: Б
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Задания закрытого типа на установление соответствия

1. Установите соответствие между термином и его определением. Каждому элементу левого столбца соответствует только один элемент правого столбца:

	Термин		Определение
1)	TF-IDF	А)	Процесс замены слов в тексте на их синонимы для изменения стиля или упрощения понимания.
2)	N-грамма	Б)	Определение правильного значения (смысла) слова в конкретном контексте, когда слово имеет несколько возможных значений.
3)	Синонимизация	В)	Метод, используемый для оценки важности слова в документе относительно коллекции документов (корпуса).
4)	Разрешение неоднозначности смысла (Word Sense Disambiguation)	Г)	Последовательность из N последовательных элементов (слов, символов и т.д.) в тексте.
5)	Семантический анали	Д)	Анализ значения слов, фраз и предложений для понимания смысла текста.

Правильный ответ: 1-В, 2-Г, 3-А, 4-Б, 5-Д
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. Установите соответствие между термином и его определением. Каждому элементу левого столбца соответствует только один элемент правого столбца:

	Термин		Определение
1)	Корпус	А)	Процесс анализа текста с целью определения его грамматической структуры.
2)	Парсинг	Б)	Метод машинного обучения, при котором алгоритму предоставляется набор данных, уже размеченных с указанием правильных ответов.
3)	Стеммер	В)	Процесс добавления информации к тексту, например, указание частей речи, выделение именованных сущностей и т.д.
4)	Разметка	Г)	Алгоритм, который приводит слова к их корневой форме путем отбрасывания

суффиксов и приставок.

5) Обучение с учителем — Д) Большой и структурированный набор текстов, используемый для обучения и оценки моделей NLP.

Правильный ответ: 1-Д, 2-А, 3-Г, 4-В, 5-Б
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. *Методы оценки качества моделей машинного обучения, используемых в NLP, и их описание. Установите соответствие методу оценки качества моделей и его описанию. Каждому элементу левого столбца соответствует только один элемент правого столбца*:

Метод		Описание
1) Кросс-валидация	А)	График, отображающий степень корректности классификации в зависимости от выбранного порогового значения
2) Матрица ошибок	Б)	Разбиение данных на n частей и последовательное использование каждой части в качестве теста, а оставшихся в качестве обучения
3) MSE / RMSE	В)	Таблица, используемая для оценки точности предсказаний бинарной классификации при несбалансированных классах
	Г)	Среднеквадратичная ошибка / корень из среднеквадратичной ошибки для оценки точности регрессионных моделей

Правильный ответ: 1-Б, 2-В, 3-Г
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4. Основные этапы построения языковой модели машинного обучения и описание. Установите соответствие этапу построения языковой модели и его описанию. Каждому элементу левого столбца соответствует только один элемент правого столбца:

Этап		Описание
1) Подготовка данных	А)	Используется для улучшения точности и обобщающей способности модели. Включает в себя анализ ошибок, настройку параметров модели, а также использование регуляризации и ансамблей моделей
2) Выбор метода и создание модели	Б)	Этап, на котором проверяется способность модели к обобщению на новые данные, не использовавшиеся в процессе обучения
3) Оценка и улучшение	В)	Этап, на котором осуществляется разметка

	качества модели		данных
4)	Оценка качества модели на тестовой выборке	Г)	Этап, на котором осуществляется сбор, очистка, преобразование данных, а также разделение на обучающую и тестовую выборки
5)		Д)	На этом этапе определяется, какой алгоритм машинного обучения будет использоваться, настраиваются параметры модели, и происходит ее обучение на обучающей выборке

Правильный ответ: 1-Г, 2-Д, 3-А, 4-Б
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Задания закрытого типа на установление правильной последовательности

1. Стандартный процесс предобработки текста для анализа. Запишите правильную последовательность букв слева направо.
А) Удаление стоп-слов.
Б) Токенизация текста.
В) Лемматизация или стемминг.
Г) Преобразование текста в нижний регистр.
Правильный ответ: Г, Б, А, В
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. Создание модели классификации текста с использованием машинного обучения. Запишите правильную последовательность букв слева направо.
А) Разделение данных на обучающую и тестовую выборки.
Б) Обучение модели машинного обучения на обучающей выборке.
В) Оценка производительности модели на тестовой выборке.
Г) Предобработка текста (токенизация, удаление стоп-слов, стемминг/лемматизация).
Д) Векторизация текста (например, TF-IDF, Word2Vec).
Правильный ответ: Г, Д, А, Б, В
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. Использование Named Entity Recognition (NER) для извлечения информации из текста. Запишите правильную последовательность букв слева направо.
А) Предобработка текста (токенизация, удаление стоп-слов и т.д.).
Б) Использование модели NER для идентификации и классификации именованных сущностей.
В) Анализ результатов NER и извлечение интересующей информации.
Правильный ответ: А, Б, В
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4. Создание модели Word Embedding (например, с использованием Word2Vec). Запишите правильную последовательность букв слева направо.
А) Обучение модели Word2Vec на большом корпусе текста.
Б) Предобработка текста (токенизация, удаление стоп-слов, и т.д.).
В) Использование обученной модели для получения векторных представлений слов.
Правильный ответ: Б, А, В
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Задания открытого типа

Задания открытого типа на дополнение

1. *Напишите пропущенное слово (словосочетание).*
Процесс разбиения текста на отдельные слова или фразы называется _____.
Правильный ответ: токенизация.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. *Напишите пропущенное слово (словосочетание).*
Задача определения эмоциональной окраски текста (позитивной, негативной или нейтральной) называется _____.
Правильный ответ: анализ тональности / sentiment analysis.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. Напишите пропущенное слово (словосочетание).
BERT, RoBERTa, и ELMo являются примерами _____ моделей, которые показывают современные результаты в различных задачах NLP.
Правильный ответ: Transformer / Трансформер
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4**.** *Напишите пропущенное слово (словосочетание).*
Набор часто встречающихся слов, которые обычно удаляются из текста при предобработке, называются _____.
Правильный ответ: стоп-слова.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Задания открытого типа с кратким свободным ответом

1. *Дайте ответ на вопрос.*
Как называется процесс разбиения текста на отдельные слова или токены?
Правильный ответ: Токенизация / Разбиение на токены / Лексический анализ
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. *Дайте ответ на вопрос.*
Как называется процесс приведения слова к его базовой или словарной форме?
Правильный ответ: Лемматизация / Приведение к лемме / Нормализация
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. *Дайте ответ на вопрос* .
Как называется метод представления текста, в котором каждое слово представляется числовым вектором, отражающим его семантическое значение?
Правильный ответ: Word embeddings / Векторные представления слов / Word2Vec
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4. *Дайте ответ на вопрос.*
Как называется задача определения эмоциональной окраски текста (например, позитивный, негативный, нейтральный)?
Правильный ответ: Анализ тональности / Определение тональности
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Задания открытого типа с развернутым ответом

1. *Дайте развернутый ответ на вопрос:*
Опишите основные шаги процесса обработки текста для анализа тональности (sentiment analysis).
Время выполнения: 20 мин.
Ожидаемый результат:
Процесс обработки текста для анализа тональности обычно включает следующие шаги:
1. Сбор данных.
2. Предобработка текста: токенизация – разделение текста на отдельные слова (токены); удаление стоп-слов – удаление часто встречающихся слов (например, “и”, “а”, “но”), которые не несут существенной смысловой нагрузки; приведение к нижнему регистру; стемминг или лемматизация – приведение слов к их корневой или словарной форме; удаление пунктуации и специальных символов.
3. Векторизация текста - преобразование текста в числовое представление, которое может быть использовано моделью машинного обучения.
4. Выбор и обучение модели: выбор алгоритма машинного обучения и обучение его на размеченных данных.
5. Оценка модели: оценка производительности модели на тестовой выборке с использованием метрик.

6. Применение модели: использование обученной модели для анализа тональности новых текстов.
Критерии оценивания: ответ должен содержать минимум четыре этапа.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

2. *Дайте развернутый ответ на вопрос:*
Объясните, как работает метод «Bag of Words» (BOW) для представления текста и каковы его ограничения.
Время выполнения: 20 мин.
Ожидаемый результат:
Метод «Bag of Words» (BOW) – это простой подход к представлению текста, при котором текст рассматривается как неупорядоченный набор слов (как «мешок слов»). Он игнорирует грамматику и порядок слов, учитывая только частоту встречаемости каждого слова в документе. Этапы:
1. Создание словаря - создается словарь всех уникальных слов во всем корпусе текстов.
2. Векторизация - для каждого документа создается вектор, где каждый элемент вектора соответствует слову из словаря. Значением элемента является частота встречаемости этого слова в документе.
Ограничения:
1. Потеря порядка слов. BOW не учитывает порядок слов, что может быть важно для понимания смысла текста.
2. Не учитывает семантику. BOW не учитывает семантические отношения между словами.
3. Размерность. Словарь может быть очень большим, особенно для больших корпусов текстов, что приводит к высокой размерности векторов и увеличению вычислительной сложности.
4. Игнорирование контекста. Не учитывается контекст, в котором употребляется слово, а это влияет на понимание.
Критерии оценивания: наличие в ответе этапов и краткое описание ограничений.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

3. *Дайте развернутый ответ на вопрос:*
Опишите основные принципы и преимущества использования Word Embedding по сравнению с TF-IDF для представления текста.
Время выполнения: 20 мин.
Ожидаемый результат:
Word Embedding - это методы представления слов в виде плотных векторов в многомерном пространстве, которые отражают семантические отношения между словами.
Основные принципы:
1. Каждое слово представляется вектором фиксированной длины.

2. Семантическая близость – слова, которые часто встречаются в схожих контекстах, имеют близкие векторы в векторном пространстве.
3. Обучение на больших объемах текста.
Преимущества Word Embedding по сравнению с TF-IDF:
1. Учет семантики – Word Embedding учитывает семантические отношения между словами, в то время как TF-IDF рассматривает слова как отдельные символы.
2. Уменьшение размерности – Word Embedding создает векторы фиксированной длины, которые обычно намного меньше, чем векторы TF-IDF
3. Лучшая производительность в задачах NLP.
4. Способность к обобщению – Word Embedding позволяют модели обобщать знания, полученные из одних текстов, на другие, даже если в этих текстах встречаются незнакомые слова.
Критерии оценивания: наличие в ответе трёх принципов и трёх преимуществ.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

4. Дайте развернутый ответ на вопрос:
Что такое стоп-слова? Зачем их удаляют при предобработке текста?
Время выполнения: 20 мин.
Ожидаемый результат:
Стоп-слова – это часто встречающиеся слова в языке, которые не несут большой смысловой нагрузки для анализа текста. Это такие слова, как предлоги, артикли, местоимения, союзы и некоторые глаголы. Они обычно присутствуют в большинстве текстов, но их удаление не сильно влияет на понимание основного смысла.
Примеры стоп-слов: “и”, “в”, “на”, “к”.
Стоп-слова удаляют при предобработке текста, чтобы:
1. Уменьшить размерность данных.
2. Улучшить результаты – удаление стоп-слов помогает сфокусироваться на наиболее важных словах, которые несут основную смысловую нагрузку, что повышает точность анализа.
3. Снизить шум – стоп-слова могут вносить шум в анализ, особенно при использовании методов, основанных на частоте встречаемости слов.
Критерии оценивания: наличие в ответе краткого определения и целей использования.
Компетенции (индикаторы): ПК-2 (ПК-2.1).

Экспертное заключение

Представленный фонд оценочных средств (далее – ФОС) по дисциплине «Технологии работы с естественным языком» соответствует требованиям ФГОС ВО.

Предлагаемые формы и средства текущего и промежуточного контроля адекватны целям и задачам реализации основной профессиональной образовательной программы по направлению подготовки 09.04.01 Информатика и вычислительная техника.

Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины представлены в полном объеме.

Виды оценочных средств, включенные в представленный фонд, отвечают основным принципам формирования ФОС.

Разработанный и представленный для экспертизы фонд оценочных средств рекомендуется к использованию в процессе подготовки обучающихся по указанному направлению 09.04.01 Информатика и вычислительная техника.

Председатель учебно-методической
комиссии Краснодонского факультета
инженерии и менеджмента (филиала) _____ Родионова О.Ю.

Лист изменений и дополнений

№ п/п	Виды дополнений и изменений	Дата и номер протокола заседания кафедры (кафедр), на котором были рассмотрены и одобрены изменения и дополнения	Подпись (с расшифровкой) заведующего кафедрой (заведующих кафедрами)