**МИНИСТЕРСТВО науки и высшего образования Российской Федерации**

федеральное Государственное бюджетное образовательное учреждение высшего образования

**«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ университет»**

|  |
| --- |
| **Утверждаю:**  \_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_» \_\_\_\_ |

**ФОНД ОЦЕНОЧНЫХ СРЕДСТВ**

**для проверки сформированности компетенций**

|  |  |
| --- | --- |
| дисциплины: | **Автоматическая обработка естественного языка** |
| направление подготовки: | **01.04.02 Прикладная математика и информатика** |
| направленность (профиль): | **Машинное обучение и анализ данных** |
| форма обучения: | **Очная, очно-заочная** |

Фонд оценочных средств рассмотрен на заседании кафедры бизнес-информатики и математики

Протокол № \_\_\_ от «\_\_\_» \_\_\_ \_\_\_\_\_\_\_

**ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ И ИНДИКАТОРОВ ИХ ДОСТИЖЕНИЯ ПО**

**УЧЕБНОЙ ДИСЦИПЛИНЕ (МОДУЛЮ)**

|  |  |
| --- | --- |
| Код и наименование компетенции | Код и наименование индикатора достижения компетенции (ИДК) |
| ПКС-2 Способен разрабатывать (совершенствовать) и внедрять новые методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными | ПКС-2.1. Способен разрабатывать методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными |
|
|
| ПКС-2.2. Способен оптимизировать и внедрять методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными в профессиональной деятельности |
|
|

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕРКИ СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ**

***Для оценки*** ПКС-2 *Способен разрабатывать (совершенствовать) и внедрять новые методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными*

**10 тестовых вопросов открытого типа по дисциплине "Автоматическая обработка естественного языка"**

1. Что такое автоматическая обработка естественного языка (NLP) и основные задачи, которые она решает?

Ответ:

Автоматическая обработка естественного языка (NLP) — это область искусственного интеллекта, занимающаяся анализом, интерпретацией и генерацией текста и речи на естественных языках. Основные задачи NLP включают:

- Токенизацию (разделение текста на слова и предложения);

- Морфологический анализ (разбор слов по морфемам);

- Распознавание именованных сущностей (NER);

- Тематическое моделирование;

- Машинный перевод;

- Синтаксический и семантический анализ;

- Генерацию текста;

- Анализ тональности и классификацию текста.

2. Что такое токенизация и почему она важна в NLP?

Ответ:

Токенизация — это процесс разбиения текста на меньшие элементы, такие как слова, предложения или символы (токены). Она важна в NLP, потому что служит первым шагом в процессе анализа текста, предоставляя основу для дальнейших этапов, таких как лемматизация, морфологический анализ и извлечение информации.

3. Объясните разницу между лемматизацией и стеммингом.

Ответ:

- Лемматизация приводит слово к его словарной форме (лемме) с учетом грамматической информации (например, "бегает" → "бегать").

- Стемминг удаляет окончание или аффиксы слов, не учитывая их грамматические свойства (например, "бегает" → "бег"). Лемматизация обычно точнее, так как использует словари и грамматические правила.

4. Что такое векторизация текста? Какие методы векторизации применяются?

Ответ:

Векторизация текста — это процесс преобразования текстовых данных в числовые представления, которые могут быть использованы алгоритмами машинного обучения. Основные методы:

- Bag of Words (BoW);

- TF-IDF (Term Frequency-Inverse Document Frequency);

- Word Embeddings (например, Word2Vec, GloVe);

- Sentence Embeddings (например, BERT, GPT).

5. Что такое распознавание именованных сущностей (Named Entity Recognition, NER)? Какие типы сущностей обычно выделяют?

Ответ:

NER — это задача, направленная на идентификацию и классификацию именованных сущностей в тексте, таких как имена людей, организации, географические названия и даты. Обычные типы сущностей включают:

- PER (персоны);

- ORG (организации);

- LOC (локации);

- DATE (даты).

6. Что такое тематическое моделирование и для каких задач оно используется?

Ответ:

Тематическое моделирование — это метод NLP для выделения скрытых тем (топиков) в большом наборе текстовых данных. Эти методы полезны для автоматического анализа больших текстовых корпусов и могут применяться, например, в исследованиях рынка, классификации документов, аналитике социальных сетей. Часто используются подходы LSA (Latent Semantic Analysis) и LDA (Latent Dirichlet Allocation).

7. Как работает анализ тональности текста? Какие подходы для этого применяются?

Ответ:

Анализ тональности (sentiment analysis) определяет эмоциональную окраску текста (например, позитивную, нейтральную или негативную). Основные подходы:

- Лексикон-ориентированные: используют заранее составленные словари с оценками тональности слов.

- Машинное обучение: используют методы классификации на размеченных данных (например, SVM, Naive Bayes).

- Глубокое обучение: используют нейронные сети и эмбеддинги для анализа сложных языковых конструкций.

8. Чем различаются синтаксический и семантический анализ текста?

Ответ:

- Синтаксический анализ (Parsing) определяет грамматическую структуру текста, анализируя правила построения предложений (например, определение подлежащего и сказуемого).

- Семантический анализ интерпретирует значение текста, работая с более абстрактными понятиями, такими как смысловое соответствие между словами или фразами.

9. Что такое модели языков (Language Models) и каково их применение?

Ответ:

Модели языков предсказывают последовательности слов в тексте или вероятность появления определенного слова, учитывая предшествующий контекст. Примеры применения:

- Машинный перевод (Google Translate);

- Генерация текста (например, ChatGPT);

- Автодополнение текста;

- Поиск информации.

Примеры языковых моделей: BERT, GPT, XLNet.

10. Какую роль играют предобученные нейронные языковые модели (например, BERT, GPT) в современном NLP?

Ответ:

Предобученные языковые модели обучаются на огромных наборах текстовых данных и способны эффективно захватывать семантические и синтаксические зависимости языка. Они дают возможность быстро адаптироваться к задачам, таким как классификация текста, NER, анализ тональности, без необходимости обучать модели "с нуля".

Вот список тестовых вопросов по дисциплине "Автоматическая обработка естественного языка" с закрытыми и комбинированными вопросами.

**Закрытые вопросы (10)**

1. Что такое токенизация?

- a) Разделение текста на предложения

- b) Разделение текста на слова

- c) Анализ синтаксической структуры

- d) Определение семантики слов

- Верный ответ: b

2. Какой алгоритм используется для стемминга?

- a) Naive Bayes

- b) Porter Stemmer

- c) K-means

- d) Decision Tree

- Верный ответ: b

3. Что такое "мешок слов"?

- a) Метод классификации

- b) Метод представления текста

- c) Алгоритм машинного обучения

- d) Система автоматического перевода

- Верный ответ: b

4. Какой из следующих языков является языком программирования для обработки естественного языка?

- a) Python

- b) HTML

- c) CSS

- d) SQL

- Верный ответ: a

5. Что такое LSTM?

- a) Алгоритм кластеризации

- b) Модель для обработки последовательностей

- c) Метод визуализации данных

- d) Статистический тест

- Верный ответ: b

6. Какой метод используется для определения тональности текста?

- a) KNN

- b) Регрессия

- c) Наивный байесовский классификатор

- d) Алгоритм генетического программирования

- Верный ответ: c

7. Какой из следующих методов является методом разметки текста?

- a) TF-IDF

- b) NER (Named Entity Recognition)

- c) PCA

- d) SVM

- Верный ответ: b

8. Что такое Word2Vec?

- a) Алгоритм для генерации текста

- b) Метод представления слов в виде векторов

- c) Инструмент для токенизации

- d) Система для автоматического перевода

- Верный ответ: b

9. Какой из следующих подходов не используется для машинного перевода?

- a) Статистический

- b) Правилосообразный

- c) Нейронный

- d) Ручной

- Верный ответ: d

10. Что такое N-grams?

- a) Метод анализа изображений

- b) Последовательности из N элементов

- c) Алгоритм для кластеризации

- d) Метод генерации меток

- Верный ответ: b

**Комбинированные вопросы (10)**

1. Выберите правильные утверждения о токенизации. (Выберите все подходящие варианты)

- a) Это процесс разделения текста на слова.

- b) Это метод анализа синтаксиса.

- c) Это может включать удаление стоп-слов.

- d) Это всегда приводит к одинаковому количеству токенов.

- Верные ответы: a, c

2. Каковы преимущества использования стемминга и лемматизации? (Выберите все подходящие варианты)

- a) Упрощение обработки текста.

- b) Увеличение точности классификации.

- c) Снижение размерности данных.

- d) Устранение многозначности слов.

- Верные ответы: a, b, c

3. Выберите методы, используемые для анализа тональности текста. (Выберите все подходящие варианты)

- a) Наивный байесовский классификатор

- b) Регрессионный анализ

- c) Логистическая регрессия

- d) K-средние

- Верные ответы: a, c

4. Какие из следующих методов используются для представления текста? (Выберите все подходящие варианты)

- a) Bag of Words

- b) TF-IDF

- c) Word2Vec

- d) PCA

- Верные ответы: a, b, c

5. Выберите правильные утверждения о NER (распознавание именованных сущностей). (Выберите все подходящие варианты)

- a) Используется для извлечения имен людей.

- b) Применяется только к английскому языку.

- c) Может использоваться для извлечения географических названий.

- d) Является частью предобработки текста.

- Верные ответы: a, c

6. Что из перечисленного относится к методам машинного перевода? (Выберите все подходящие варианты)

- a) Статистический машинный перевод

- b) Нейронный машинный перевод

- c) Правилосообразный перевод

- d) Грамматический анализ

- Верные ответы: a, b, c

7. Выберите правильные утверждения о Word2Vec. (Выберите все подходящие варианты)

- a) Это метод для представления слов в виде векторов.

- b) Он основан на статистическом анализе.

- c) Он не учитывает контекст слов.

- d) Позволяет выявлять семантические связи между словами.

- Верные ответы: a, b, d

8. Какие из следующих алгоритмов могут использоваться для кластеризации текстов? (Выберите все подходящие варианты)

- a) K-means

- b) DBSCAN

- c) Наивный байесовский классификатор

- d) Иерархическая кластеризация

- Верные ответы: a, b, d

9. Что из перечисленного является частью процесса обработки естественного языка? (Выберите все подходящие варианты)

- a) Токенизация

- b) Лемматизация

- c) Классификация

- d) Сортировка данных

- Верные ответы: a, b, c

10. Выберите методы, которые могут использоваться для построения языковых моделей. (Выберите все подходящие варианты)

- a) N-grams

- b) LSTM

- c) KNN

- d) Decision Tree

- Верные ответы: a, b

**Критерии оценивания уровня сформированности компетенций**

Оценка «отлично» выставляется обучающемуся, если им даны правильные ответы более, чем на 91% вопросов.

Оценка «хорошо» выставляется обучающемуся, если им даны правильные ответы на 76-90% вопросов.

Оценка «удовлетворительно» выставляется обучающемуся, если им даны правильные ответы на 61-75% вопросов.

Оценка «неудовлетворительно» выставляется обучающемуся, если им даны правильные ответы менее чем на 60% вопросов.