

**Комплект оценочных материалов по дисциплине
«Обработка естественного языка»**

Задания закрытого типа

Задания закрытого типа на выбор правильного ответа

1. Выберите один правильный ответ

Какой из перечисленных инструментов используется для обработки естественного языка?

- A) TensorFlow
- Б) Apache Hadoop
- В) MySQL
- Г) Microsoft Excel

Правильный ответ: А

Компетенции (индикаторы): ПК-2

2. Выберите один правильный ответ

Какой из перечисленных методов используется для извлечения именованных сущностей (NER)?

- A) Метод опорных векторов (SVM)
- Б) Алгоритм CRF (Conditional Random Fields)
- В) Метод k-средних (k-means)
- Г) Алгоритм DBSCAN

Правильный ответ: Б

Компетенции (индикаторы): ПК-2

3. Выберите один правильный ответ

Какой из перечисленных методов используется для лемматизации текста?

- A) Метод опорных векторов (SVM)
- Б) Алгоритм стемминга Портера
- В) Метод главных компонент (PCA)
- Г) Алгоритм k-ближайших соседей (k-NN)

Правильный ответ: Б

Компетенции (индикаторы): ПК-2

4. Выберите один правильный ответ

Какой из следующих форматов данных используется для хранения текстовых корпусов?

- A) JSON
- Б) CSV
- В) XML
- Г) Все перечисленные

Правильный ответ: Г

Компетенции (индикаторы): ПК-2

Задания закрытого типа на установление соответствия

1. Установите соответствие между методами и их описаниями:

Метод	Описание
1) Лемматизация	А) Метод представления слов в виде векторов в векторном пространстве
2) Стемминг	Б) Метод оценки важности слова в документе относительно коллекции документов
3) TF-IDF	В) Упрощение слова до его основы (например, «бежал» → «беж»)
4) Word2Vec	Г) Приведение слова к его начальной форме (например, «бежал» → «бежать»)

Правильный ответ:

1	2	3	4
Г	В	В	А

Компетенции (индикаторы): ПК-2

2. Установите соответствие между задачами NLP и методами их решения:

Задача NLP	Метод решения
1) Классификация текста	А) Conditional Random Fields (CRF)
2) Анализ тональности	Б) Логистическая регрессия
3) Извлечение сущностей	В) Рекуррентные нейронные сети (RNN)
4) Генерация текста	Г) Метод опорных векторов (SVM)

Правильный ответ:

1	2	3	4
Б	Г	А	В

Компетенции (индикаторы): ПК-2

3. Простой текстовый файл без структуры 4. Установите соответствие между форматами данных и их описанием:

Формат данных	Описание
1) JSON	А) Простой текстовый файл без структуры
2) CSV	Б) Расширяемый язык разметки для хранения структурированных данных
3) XML	В) Табличный формат данных, разделенный запятыми
4) TXT	Г) Текстовый формат для хранения структурированных данных

Правильный ответ:

1	2	3	4
---	---	---	---

Г	В	Б	А
---	---	---	---

Компетенции (индикаторы): ПК-2

4. Установите соответствие между этапами обработки текста и их описанием:

Этап обработки текста	Описание
1) Токенизация	А) Удаление часто встречающихся, но малозначимых слов
2) Удаление стоп-слов	Б) Преобразование текста в числовое представление
3) Нормализация	В) Разбиение текста на отдельные слова или токены
4) Векторизация	Г) Приведение текста к единому формату (например, нижний регистр)

Правильный ответ:

1	2	3	4
В	А	Г	Б

Компетенции (индикаторы): ПК-2

Задания закрытого типа на установление правильной последовательности

Установите правильную последовательность.

Запишите правильную последовательность букв слева направо.

1. Установите правильную последовательность этапов предобработки текста:

- А) Лемматизация
- Б) Удаление стоп-слов
- В) Токенизация
- Г) Векторизация

Правильная последовательность: В, Б, А, Г

Компетенции (индикаторы): ПК-2

2. Установите правильную последовательность этапов работы с текстовыми данными в NLP:

- А) Оценка качества модели
- Б) Векторизация текста
- В) Предобработка текста (токенизация, лемматизация и т.д.)
- Г) Сбор и очистка данных
- Д) Обучение модели

Правильная последовательность: Г, В, Б, Д, А

Компетенции (индикаторы): ПК-2

3. Установите правильную последовательность этапов создания языковой модели:

- А) Оценка качества модели
- Б) Подготовка текстового корпуса
- В) Обучение модели (например, Word2Vec или BERT)
- Г) Токенизация и предобработка текста
- Д) Применение модели для решения задач (например, классификация или генерация текста)

Правильная последовательность: Б, Г, В, А, Д

Компетенции (индикаторы): ПК-2

4. Установите правильную последовательность этапов извлечения именованных сущностей (NER):

- А) Разметка текста (например, с использованием BIO-разметки)
- Б) Применение модели для извлечения сущностей из нового текста
- В) Токенизация текста
- Г) Оценка точности извлечения
- Д) Обучение модели (например, CRF или BiLSTM)

Правильная последовательность: В, А, Д, Б, Г

Компетенции (индикаторы): ПК-2

Задания открытого типа

Задания открытого типа на дополнение

Напишите пропущенное слово (словосочетание).

1. _____ – метод, который преобразует слово в его начальную форму.

Правильный ответ: лемматизация

Компетенции (индикаторы): ПК-2

2. _____ – метод оценки важности слова в документе относительно коллекции документов.

Правильный ответ: TF-IDF

Компетенции (индикаторы): ПК-2

3. _____ – метод, который используется для преобразования текста в числовое представление.

Правильный ответ: векторизация

Компетенции (индикаторы): ПК-2

4. _____ – метод, который используется для разбиения текста на отдельные слова или токены.

Правильный ответ: токенизация
Компетенции (индикаторы): ПК-2

Задания открытого типа с кратким свободным ответом

Напишите пропущенное слово (словосочетание).

1. К основным этапам предобработки текста относится: токенизация, удаление стоп-слов, лемматизация или стемминг, _____ (приведение к нижнему регистру), векторизация.

Правильный ответ: нормализация
Компетенции (индикаторы): ПК-2

2. Задача NER (Named Entity Recognition) – это задача извлечения _____ из текста. Обычно извлекаются такие типы сущностей, как имена людей, организации, локации, даты и др.

Правильный ответ: именованных сущностей
Компетенции (индикаторы): ПК-2

3. _____ – это часто встречающиеся, но малозначимые слова (например, «и», «в», «на»). Их удаляют, чтобы уменьшить размер данных и сосредоточиться на более важных словах.

Правильный ответ: Стоп-слова
Компетенции (индикаторы): ПК-2

4. Токенизация – это процесс _____ на отдельные слова или токены. Она нужна для дальнейшей обработки текста, например, для лемматизации, векторизации или анализа.

Правильный ответ: разбиения текста
Компетенции (индикаторы): ПК-2

Задания открытого типа с развернутым ответом

1. Опишите основные этапы предобработки текста и объясните, зачем каждый из них нужен.

Время выполнения – 15 мин.

Критерии оценивания: полное содержательное соответствие приведенному ниже пояснению

Ожидаемый результат:

Основные этапы предобработки текста:

1. Токенизация – разбиение текста на слова или предложения. Нужна для дальнейшей обработки текста на уровне отдельных элементов.

2. Удаление стоп-слов – удаление часто встречающихся, но малозначимых слов (например, «и», «в», «на»). Это уменьшает размер данных и помогает сосредоточиться на важных словах.

3. Лемматизация или стемминг – приведение слов к их начальной форме (лемматизация) или основе (стемминг). Это помогает унифицировать слова и уменьшить размерность данных.

4. Нормализация – приведение текста к единому формату (например, нижний регистр). Это упрощает анализ и сравнение слов.

5. Векторизация – преобразование текста в числовое представление (например, с использованием TF-IDF или Word2Vec). Это необходимо для использования текстовых данных в машинном обучении.

Компетенции (индикаторы): ПК-2

2. Опишите, как работает метод анализа тональности текста, и приведите пример его применения.

Время выполнения – 15 мин.

Критерии оценивания: полное содержательное соответствие приведенному ниже пояснению

Ожидаемый результат:

Анализ тональности текста – это задача классификации текста на положительный, отрицательный или нейтральный.

Этапы:

1. Предобработка текста (токенизация, удаление стоп-слов, лемматизация).

2. Векторизация текста (например, с использованием TF-IDF или Word2Vec).

3. Обучение модели классификации (например, SVM, логистическая регрессия или нейронные сети).

4. Оценка качества модели на тестовых данных.

Пример применения: анализ отзывов пользователей, мониторинг мнений в социальных сетях, оценка тональности новостей.

Компетенции (индикаторы): ПК-2

3. Вам нужно создать чат-бот для ответов на вопросы пользователей. Опишите, как вы будете подходить к этой задаче, и какие модели или инструменты вы будете использовать.

Время выполнения – 15-20 мин.

Критерии оценивания: содержательное соответствие приведенному ниже пояснению

Ожидаемый результат:

1. Сбор данных:

- Создание базы вопросов и ответов.
- Разметка данных для обучения модели.

2. Построение модели:

- Использование моделей, таких как Seq2Seq, GPT или BERT.
- Обучение модели на диалоговых данных.

3. Оценка качества:

- Тестирование чат-бота на реальных вопросах.
- Анализ ошибок и улучшение модели.

Компетенции (индикаторы): ПК-2

4. Напишите фрагмент кода на Python с использованием библиотеки spaCy для извлечения именованных сущностей (NER) из следующего текста: "Apple is looking at buying U.K. startup for \$1 billion.". Выведите текст и соответствующие сущности.

Время выполнения – 20 мин.

Критерии оценивания: содержательное соответствие приведенному ниже программному коду и пояснениям к нему

Ожидаемый результат:

```
import spacy
# Загрузка модели spaCy
nlp = spacy.load("en_core_web_sm")
# Пример текста
text = "Apple is looking at buying U.K. startup for $1 billion."
# Обработка текста
doc = nlp(text)
# Извлечение и вывод именованных сущностей
for ent in doc.ents:
    print(f"Текст: {ent.text}, Сущность: {ent.label_}")
```

Пояснение:

1. Загрузка модели: Используем предобученную модель en_core_web_sm.
2. Обработка текста: Применяем модель к тексту.
3. Извлечение сущностей: Выводим текст и тип сущности (например, ORG, GPE, MONEY).

Компетенции (индикаторы): ПК-2

Экспертное заключение

Представленный комплект оценочных материалов по дисциплине «Обработка естественного языка» соответствует требованиям ФГОС ВО.

Предлагаемые оценочные материалы адекватны целям и задачам реализации основной профессиональной образовательной программы по направлению подготовки 09.03.03 Прикладная информатика.

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

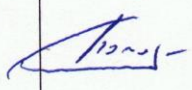
Разработанные и представленные для экспертизы оценочные материалы рекомендуются к использованию в процессе подготовки обучающихся по указанному направлению.

Председатель учебно-методической комиссии
института компьютерных систем
и информационных технологий



Ветрова Н.Н.

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)
1.	Дополнен комплект оценочных материалов	протокол заседания кафедры компьютерных систем и сетей № <u>8</u> от <u>10.03.2025</u>	 С.В. Попов