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

Введение в анализ естественного языка (NLP)

по направлению подготовки

45.04.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
Компьютерная и когнитивная лингвистика

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

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-3 Способен выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий.

ОПК-4 Способен расширять сферу научной деятельности, участвовать в междисциплинарных исследованиях на стыке наук.

ОПК-6 Способен осуществлять эффективное управление разработкой программных средств информационных проектов в сфере своей профессиональной деятельности.

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

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-3.2 Критически сопоставляет и оценивает существующие подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий

ИОПК-4.1 Демонстрирует знание новых теорий в сфере междисциплинарного взаимодействия лингвистики и наук гуманитарного, математического и естественно-научного циклов

ИОПК-6.1 Аргументированно выбирает математические и лингвистические методы решения профессиональных задач с применением языков программирования

ИПК-1.1 Обнаруживает знания об актуальных направлениях междисциплинарных лингвистических исследований в избранной научной сфере

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-4.2 Разрабатывает программу действий по решению задач проекта в области когнитивной и компьютерной лингвистики с учетом имеющихся технических средств и информационных технологий, в том числе в области искусственного интеллекта

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

Элементы текущего контроля:
– тесты;
– контрольная работа;

Тест (ИОПК-3.2.)

1. Каково назначение эмбеддингов слов в NLP?
 а) представить слова в виде плотных числовых векторов
 б) для анализа настроения, выраженного в тексте
 в) для выполнения тегирования частей речи
 г) генерировать связные и осмысленные предложения
2. Какой из следующих алгоритмов обычно используется для распознавания именованных сущностей в NLP?
 а) машины опорных векторов (SVM)
 б) классификатор Наивного Байеса
 в) длительная кратковременная память (LSTM)

г) кластеризация K-средних
3. Как называется процесс присвоения синтаксической структуры предложению?
а) стемминг
б) токенизация
в) синтаксический разбор
г) лемматизация
Ключи:
1. а
2. а
3. в

Тест (ИОПК-4.1)
1. Какая техника используется для решения проблемы полисемии в NLP?
а) распознавание именованных сущностей
b) разрешение кореференции
с) разотождествление смысла слов
г) сентимент-анализ
2. Что из нижеперечисленного является примером модели "последовательность-последовательность" в NLP?
а) конволюционная нейронная сеть (CNN)
b) рекуррентная нейронная сеть (RNN)
в) машины опорных векторов (SVM)
г) дерево решений
3. Какова цель тематического моделирования в NLP?
а) классифицировать документы по заранее определенным категориям
б) анализ настроения, выраженного в тексте
с) генерировать связные и осмысленные предложения
г) выполнение тегирования частей речи
Ключи:
1. в
2. б
3. а

Тест (ИОПК-6.1)
1. Какой алгоритм обычно используется для машинного перевода в NLP?
а) случайный лес
b) K-nearest neighbors (KNN)
в) трансформер
г) дерево решений
2. Как называется процесс определения настроения, выраженного в тексте?
а) Анализ настроения
б) Классификация текстов
в) Распознавание именованных сущностей
г) Топологическое моделирование

Ключи:
1. в
2. а

Тест (ИПК-1.1)
1. Какая техника используется для решения проблемы нехватки данных в языковом моделировании?

а) N-граммы
б) эмбеддинги слов
в) Латентное распределение Дирихле (LDA)
г) иерархические сети внимания (HAN)

2. Какова цель разрешения кореференции в NLP?
a) присвоение грамматических меток словам в предложении
b) сгруппировать слова с похожими значениями вместе
c) разрешить ссылки на ранее упомянутые сущности
г) анализ синтаксической структуры предложения

Ключи:
1. а
2. в

Критерии оценивания: тест считается пройденным, если обучающий ответил правильно как минимум на половину вопросов.

Контрольная работа (ИПК-4.1, ИПК-4.2)

Контрольная работа состоит из 2 задач.

Примеры задач:

Задача 1 (ИПК-4.1)

Подключитесь к OpenAI API и задайте генеративной системе роль помощника инструктура с помощью промпт-инжиниринга.

Ответ

```
from dotenv import load_dotenv
load_dotenv()
import openai
# For GPT 3.5 Turbo, the endpoint is ChatCompletion
response = openai.ChatCompletion.create(
   # For GPT 3.5 Turbo, the model is “gpt-3.5-turbo”
  model="gpt-3.5-turbo",
   # Conversation as a list of messages.
   messages=[
      {"role": "system", "content": "You are a helpful teacher."},
      {
         "role": "user",
         "content": "Is there other measures than time complexity for an \
         algorithm?",
      },
      {
         "role": "assistant",
         "content": "Yes, there are other measures besides time complexity \
         for an algorithm, such as space complexity.",
      },
      {"role": "user", "content": "What is it?"},
   ],
)
print (response["choices"][0]["message"]["content"])
```

Задача 2 (ИПК-4.2)

Проверьте Ваш текст с помощью модели-модератора. Напишите код и сохраните вывод.

Ответ:

```
import openai
# Call the openai Moderation endpoint, with the text-moderation-latest model
response = openai.Moderation.create(
model="text-moderation-latest",
input="I want to kill my neighbor.",
)
```

Пример вывода:


```
{
"id": "modr-7AftIJg7L5jqGIsbc7NutObH4j0Ig",
"model": "text-moderation-004",
"results": [
{
"categories": {
"hate": false,
"hate/threatening": false,
"self-harm": false,
"sexual": false,
"sexual/minors": false,
"violence": true,
"violence/graphic": false,
},
"category_scores": {
"hate": 0.0400671623647213,
"hate/threatening": 3.671687863970874e-06,
"self-harm": 1.3143378509994363e-06,
"sexual": 5.508050548996835e-07,
"sexual/minors": 1.1862029225540027e-07,
"violence": 0.9461417198181152,
"violence/graphic": 1.463699845771771e-06,
},
"flagged": true,
}
],
}
```


Критерии оценивания:

Результаты контрольной работы определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

При оценивании ответа главное внимание уделяется демонстрации обучающимся знаний о написании программного кода, специфики библиотек и синтаксиса языка Python.

За каждый вопрос и практическое задание выставляется отдельная оценка, общая экзаменационная оценка складывается из трех частных оценок. Если студент получает оценку «неудовлетворительно» за один из вопросов или за практическое задание, общая положительная оценка не выставляется.

Оценка	Критерии	
	Теоретический вопрос	Практическое задание
Отлично	Ответ полный, проиллюстрированный самостоятельно подобранными примерами, содержание структурировано, логика ответа прозрачна.	Задание выполнено без ошибок или допущена 1 незначительная ошибка
Хорошо	А) Ответ полный, но не структурированный, примеры	Задание выполнено с 2 ошибками, которые

	заимствованы из разобранных на занятиях или некачественно проинтерпретированы. Б) В теоретической части ответа имеются отдельные лакуны, которые могут быть заполнены на основании дополнительных вопросов.	студент способен исправить после указания на нее.
Удовлетворительно	Ответ не полный, слабо структурированный, студент некачественно устанавливает связи обсуждаемой проблемы с другими положениями изученного курса, допускает грубые ошибки в интерпретации примеров.	Задание выполнено с 3 ошибками.
Неудовлетворительно	При ответе допускаются грубые теоретические ошибки, обнаруживаются пробелы в знаниях важнейших теоретических положений курса, студент не способен приводить и интерпретировать примеры.	Задание не выполнено либо при его выполнении допущено более 4 ошибок.

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

Первый семестр, экзамен

Экзаменационный билет состоит из двух частей.

Первая часть представляет собой батарею тестов из вопросов, проверяющих набор компетенций. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Тест (ИОПК-3.2)

1. Что такое обработка естественного языка (NLP)?

а) изучение того, как люди овладевают языком

б) изучение того, как компьютеры понимают и обрабатывают человеческий язык

в) изучение того, как языки развиваются с течением времени

г) изучение того, как язык используется в социальных контекстах.

2. Какая из перечисленных ниже задач НЕ является распространенным применением NLP?

а) машинный перевод

б) анализ настроения

в) распознавание изображений

г) распознавание именованных сущностей

3. Как называется процесс разбиения предложения на грамматические составляющие?

а) токенизация

б) стемминг

в) лемматизация

г) синтаксический разбор

Ключи:

1. б
2. в
3. г

Тест (ИОПК-4.1)

1. Что из перечисленного ниже является примером техники синтаксического разбора?

а) тегирование частей речи

b) распознавание именованных сущностей

в) анализ настроения

г) топологическое моделирование

2. Каково назначение вкраплений слов в NLP?

а) представить слова в виде плотных числовых векторов

б) удаление из текста стоп-слов

с) идентифицировать именованные сущности в предложении

г) классифицировать документы по заранее определенным категориям

3. Какой из следующих алгоритмов обычно используется для решения задач маркировки последовательностей в NLP?

а) наивный Байес

б) машины опорных векторов (SVM)

в) скрытые марковские модели (HMM)

г) кластеризация K-средних

Ключи:

1. а
2. а
3. в

Тест (ИОПК-6.1)

1. Как называется процесс определения настроения или эмоций, выраженных в тексте?

а) анализ настроения

б) классификация текстов

в) топологическое моделирование

г) распознавание именованных сущностей

2. Что из перечисленного ниже является примером задачи генерации языка в NLP?

а) резюме текста

б) распознавание именованных сущностей

с) тегирование частей речи

г) анализ настроения

Ключи:

1. а
2. а

Тест (ИПК-1.1)

1. Какова цель распознавания именованных сущностей в NLP?

а) определить настроение, выраженное в тексте

б) классифицировать документы по заранее определенным категориям

с) извлечение и классификация именованных сущностей, таких как имена, местоположения и организации

г) создание связных и осмысленных предложений

2. Что из перечисленного ниже НЕ является проблемой при обработке естественного языка?

а) неоднозначность языка

б) отсутствие маркированных обучающих данных
в) высокая точность и эффективность алгоритмов
г) языковые вариации и диалекты

Ключи:
1. в
2. в

Вторая часть содержит задачи на программирование, проверяющее ИПК-4.1. и ИПК-4.2

Задача 1 (ИПК-4.1)
Как отправить текстовую последовательность на вход ChatGPT-3.5 Turbo?

Ответ:
1. Install the openai dependency:
pip install openai
2. Set your API key as an environment variable:
export OPENAI_API_KEY=sk-(...)
3. In Python, import openai:
import openai
4. Call the openai.ChatCompletion endpoint:
response = openai.ChatCompletion.create(
model="gpt-3.5-turbo",
messages=[{"role": "user", "content": "Your Input Here"}],
)
5. Get the answer:
print(response['choices'][0]['message']['content'])
Задача 2 (ИПК-4.2)
Как превратить PDF-файл в эмбеддинг с помощью ChatGPT? Напишите функцию.

Ответ:
def pdf_to_embeddings(self, pdf_path: str, chunk_length: int = 1000):
# Read data from pdf file and split it into chunks
reader = PdfReader(pdf_path)
chunks = []
for page in reader.pages:
text_page = page.extract_text()
chunks.extend([text_page[i:i+chunk_length]
for i in range(0, len(text_page), chunk_length)])
# Create embeddings
response = openai.Embedding.create(model='text-embedding-ada-002',
input=chunks)
return [{'id': value['index'],
'vector':value['embedding'],
'text':chunks[value['index']]} for value]

Критерии оценивания:
Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

При оценивании ответа на зачете с оценкой главное внимание уделяется демонстрации обучающимся знаний об особенностях работы с библиотекой OpenAI и написании эффективных запросов к современным API в области NLP.

За каждый вопрос и практическое задание выставляется отдельная оценка, общая экзаменационная оценка складывается из трех частных оценок. Если студент получает оценку «неудовлетворительно» за один из вопросов или за практическое задание, общая положительная оценка не выставляется.

Оценка	Критерии	
	Теоретический вопрос	Практическое задание
Отлично	Ответ полный, проиллюстрированный самостоятельно подобранными примерами, содержание структурировано, логика ответа прозрачна.	Задание выполнено без ошибок или допущена 1 незначительная ошибка
Хорошо	А) Ответ полный, но не структурированный, примеры заимствованы из разобранных на занятиях или некачественно проинтерпретированы. Б) В теоретической части ответа имеются отдельные лакуны, которые могут быть заполнены на основании дополнительных вопросов.	Задание выполнено с 2 ошибками, которые студент способен исправить после указания на нее.
Удовлетворительно	Ответ не полный, слабо структурированный, студент некачественно устанавливает связи обсуждаемой проблемы с другими положениями изученного курса, допускает грубые ошибки в интерпретации примеров.	Задание выполнено с 3 ошибками.
Неудовлетворительно	При ответе допускаются грубые теоретические ошибки, обнаруживаются пробелы в знаниях важнейших теоретических положений курса, студент не способен приводить и интерпретировать примеры.	Задание не выполнено либо при его выполнении допущено более 4 ошибок.

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

5 заданий, проверяющих ИОПК-3.2 Критически сопоставляет и оценивает существующие подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий.

Задание 1

1. Какова основная цель обработки естественного языка (NLP)?
а) Разработка компьютерных программ, способных понимать и генерировать человеческий язык
б) Изучение эволюции и истории человеческих языков

с) Анализ культурных и социальных аспектов использования языка
г) Изучение когнитивных процессов, связанных с усвоением языка.

2. Что из перечисленного ниже НЕ является подобластью НЛП?
а) Распознавание речи
б) Анализ чувств
в) Компьютерное зрение
г) Распознавание именованных сущностей

3. Как называется процесс преобразования устной речи в письменный текст?
а) Синтез речи
б) Распознавание речи
в) Резюме текста
г) анализ настроения

4. Что из перечисленного ниже является примером языковой модели, используемой в НЛП?
а) Классификатор Наивного Байеса
б) Машина опорных векторов (SVM)
в) Длительная кратковременная память (LSTM)
г) кластеризация K-средних

5. Какова цель использования тегов частей речи в НЛП?
а) Определение настроения, выраженного в тексте
b) Классифицировать документы по заранее определенным категориям
с) Присвоение грамматических меток словам в предложении
г) Создание связных и осмысленных предложений

Ключи:
1. а
2. в
3. б
4. в
5. в

Задание 2

1. В чем заключается одна из ключевых проблем коммерциализации машинного обучения?
а) Недостаток доступных данных
б) Ограниченные вычислительные мощности
в) Высокая стоимость оборудования
г) Невозможность эффективного обучения моделей

2. Что из перечисленного ниже НЕ является распространенным применением машинного обучения в бизнесе?
а) Обнаружение мошенничества
б) Сегментация клиентов
в) Анализ настроений
г) Управление цепочками поставок

3. Как называется процесс преобразования необработанных данных в формат, подходящий для алгоритмов машинного обучения?
а) Очистка данных
б) Предварительная обработка данных
в) Разработка признаков
г) Обучение модели

4. Какой из перечисленных ниже фреймворков машинного обучения используется в коммерческих приложениях?
а) TensorFlow
б) PyTorch
в) Keras
г) Scikit-learn

5. Какова цель оценки моделей в коммерческом машинном обучении?
а) Оценить производительность обученных моделей
б) Оптимизировать гиперпараметры
в) Для предварительной обработки данных для обучения
г) Визуализация результатов предсказаний

Ключи:
1. а
2. г
3. б
4. а
5. а

Задание 3
Токенизируйте произвольный текст, используя библиотеку spacy. Напишите соответствующий код

Ответ
```
import spacy
nlp = spacy.load("en_core_web_md")
doc = nlp("I own a ginger cat.")
print ([token.text for token in doc])
['I', 'own', 'a', 'ginger', 'cat', '.']
```

Задание 4
Приведите пример кастомизации лемматизатора spacy, используя произвольную последовательность, не включенную в словарь по умолчанию.

Ответ
```
import spacy
from spacy.symbols import ORTH
nlp = spacy.load("en_core_web_md")
doc = nlp("lemme that")
print([w.text for w in doc])
['lemme', 'that']
special_case = [{ORTH: "lem"}, {ORTH: "me"}]
nlp.tokenizer.add_special_case("lemme", special_case)
print([w.text for w in nlp("lemme that")])
```

['lem', 'me', 'that']

Задание 5
Используйте spacy для объяснения используемых в библиотеке понятий

Ответ
spacy.explain("NNS)
'noun, plural'
doc = nlp("I saw flowers.")
token = doc[2]
token.text, token.tag_, spacy.explain(token.tag_)
('flowers', 'NNS', 'noun, plural')

5 заданий, проверяющих ИОПК-4.1 Демонстрирует знание новых теорий в сфере междисциплинарного взаимодействия лингвистики и наук гуманитарного, математического и естественно-научного циклов.

Задание 1

1. Какой алгоритм обычно используется для машинного перевода в НЛП?
   а) Дерево решений
   б) Случайный лес
   в) Рекуррентная нейронная сеть (RNN)
   г) K-nearest neighbors (KNN)

2. Как называется процесс извлечения ключевой информации из текста?
   а) Распознавание именованных сущностей
   б) Анализ настроения
   в) Классификация текста
   г) Резюме текста

3. Что из перечисленного ниже является примером задачи классификации текста в НЛП?
   а) Анализ настроения
   б) Распознавание именованных сущностей
   в) Машинный перевод
   г) Синтез речи

4. Какова цель синтаксического анализа зависимостей в НЛП?
   а) Выявление связей между словами в предложении
   б) Классифицировать документы по заранее определенным категориям
   с) Генерировать связные и осмысленные предложения
   d) Извлечение и классификация именованных сущностей, таких как имена, местоположения и организации.

5. Для чего в НЛП используются n-граммы?
   а) Для представления слов в виде плотных числовых векторов
   б) Для анализа настроения, выраженного в тексте
   в) Для моделирования вероятности последовательности слов
   г) Для классификации документов по заранее определенным категориям

1. в
2. а

3. а
4. а
5. в

Задание 2

1. Что из перечисленного ниже является распространенным методом развертывания моделей машинного обучения в производстве?
   а) Облачные платформы
   б) Локальные серверы
   с) Распределенные системы
   г) Виртуальные машины

2. Что является одним из этических соображений при коммерциализации машинного обучения?
   а) Предвзятость данных и алгоритмов
   б) Чрезмерная подгонка моделей
   в) Сложность алгоритмов
   г) Отсутствие возможности интерпретации

3. Что из перечисленного ниже является примером неконтролируемого обучения в коммерческом машинном обучении?
   а) Классификация изображений
   б) Анализ настроения
   в) Обнаружение аномалий
   г) Регрессионный анализ

4. Как называется процесс тонкой настройки предварительно обученной модели машинного обучения на конкретных данных?
   а) Трансферное обучение
   б) Обучение с усилением
   в) ансамблевое обучение
   г) активное обучение

5. Что является одним из ключевых моментов для масштабирования решений машинного обучения в коммерческих условиях?
   а) Эффективное хранение данных
   б) Прогнозирование в реальном времени
   в) Выбор алгоритма
   г) Интерпретируемость моделей

1. а
2. а
3. в
4. а
5. б

Задание 3
С помощью какого скрипта можно получить следующий фрагмент вывода?
('Alicia', 'PROPN', 'NNP', 'proper noun', 'noun, proper singular')
('and', 'CCONJ', 'CC', 'coordinating conjunction', 'conjunction,
coordinating')

('me', 'PRON', 'PRP', 'pronoun', 'pronoun, personal')
('went', 'VERB', 'VBD', 'verb', 'verb, past tense')
('to', 'ADP', 'IN', 'adposition', 'conjunction, subordinating or
preposition')

Ответ:
```
import spacy
nlp = spacy.load("en_core_web_md")
doc = nlp("Alicia and me went to the school by bus.")
for token in doc:
token.text, token.pos_, token.tag_, \
spacy.explain(token.pos_), spacy.explain(token.tag_)
```

Задание 4
Напишите пример кода, проверяющего способность spacy различать омонимы

Ответ
```
doc = nlp("I will ship the package tomorrow.")
for token in doc:
token.text, token.tag_, spacy.explain(token.tag_)

doc = nlp("I saw a red ship.")
for token in doc:
token.text, token.tag_, spacy.explain(token.tag_)
```

Задание 5
Что означают следующие синтаксические теги spacy?

1. amod:
2. aux:
3. compound:
4. dative:
5. det:
6. dobj:
7. nsubj:
8. nsubjpass:
9. nummod:
10. poss:
11. root:

Ответ:

12. Adjectival modifier
13. Auxiliary
14. Compound
15. Dative object
16. Determiner
17. Direct object
18. Nominal subject
19. Nominal subject, passive
20. Numeric modifier
21. Possessive modifier
22. The root

5 заданий, проверяющих ИОПК-6.1 Аргументированно выбирает математические и лингвистические методы решения профессиональных задач с применением языков программирования.

Задание 1

1. Какова основная цель обработки естественного языка (NLP)?
   а) Изучить происхождение языков
   б) Разработка компьютерных систем, способных понимать и обрабатывать человеческий язык
   в) Проанализировать культурное влияние языка
   г) Изучение психологических аспектов овладения языком.

2. Что из перечисленного ниже НЕ является распространенной задачей НЛП?
   а) Анализ смыслового содержания
   б) Распознавание именованных сущностей
   в) Распознавание образов
   г) Машинный перевод

3. Как называется процесс преобразования устной речи в письменный текст?
   а) Распознавание речи
   б) Синтез речи
   в) Анализ настроения
   г) Резюме текста

4. Какой алгоритм обычно используется для тегирования частей речи в НЛП?
   а) Деревья решений
   б) Машины опорных векторов (SVM)
   с) Скрытые марковские модели (HMM)
   г) кластеризация K-средних

5. С какой целью в НЛП используется стемминг?
   а) Для определения настроения, выраженного в тексте
   b) Удаление из текста стоп-слов
   с) Сгруппировать слова с похожими значениями вместе
   г) Анализ синтаксической структуры предложения

Ключи:
1. б
2. в
3. а
4. в
5. в

Задание 2

1. Что такое чатбот?
   а) Физический робот, взаимодействующий с людьми
   б) Программное обеспечение, имитирующее человеческий разговор
   в) Устройство с искусственным интеллектом, выполняющее бытовые задачи
   г) Виртуальный помощник, помогающий решать административные задачи.

2. Какая из перечисленных ниже технологий обычно используется при разработке чатботов?
а) Обработка естественного языка (NLP)
б) дополненная реальность (AR)
с) Виртуальная реальность (VR)
г) блокчейн

3. Какова цель распознавания намерений в чат-ботах?
а) Понять эмоции пользователя
б) Определить местоположение пользователя
с) Определить желаемое действие или цель пользователя
г) Анализ мимики пользователя.

4. Какой из следующих типов чатботов использует предопределенные правила и скрипты для ответа на запросы пользователей?
а) Чатботы, основанные на правилах
б) чатботы на основе искусственного интеллекта
в) чатботы с машинным обучением
г) гибридные чатботы

5. Как называется процесс обучения чатбота на наборе данных разговоров?
а) предварительная обработка чатбота
б) Расширение чатбота
с) Тонкая настройка чатбота
г) нормализация чатбота

Ключи:
1. б
2. а
3. в
4. а
5. в

Задание 3
Напишите код, использующий и объясняющий NER-функционал библиотеки spacy

Ответ
doc = nlp("Albert Einstein was born in Ulm on 1879. He studied
electronical engineering at ETH Zurich.")
doc.ents
(Albert Einstein, Ulm, 1879, ETH Zurich)
for token in doc:
token.text, token.ent_type_, \
spacy.explain(token.ent_type_)

Задание 4
Напишите код, извлекающий номера телефонов из текста. Формат номеров произвольный. Используйте библиотеку spacy.

Ответ
import spacy

```
from spacy.matcher import Matcher

nlp = spacy.load("en_core_web_md")

doc1 = nlp("You can call my office on +1 (221) 102-2423 or email me directly.")
doc2 = nlp("You can call me on (221) 102 2423 or text me.")

pattern = [{"TEXT": "+1", "OP": "?"}, {"TEXT": "("}, {"SHAPE": "ddd"}, {"TEXT": ")"},
{"SHAPE": "ddd"}, {"TEXT": "-", "OP": "?"}, {"SHAPE": "dddd"}]

matcher = Matcher(nlp.vocab)
matcher.add("usPhonNum", [pattern])

for mid, start, end in matcher(doc1):
    print(start, end, doc1[start:end])
```

Задание 5
Напишите код, выбирающий несколько имен собственных из текста, используя Matcher библиотеки spacy.

Ответ
```
import spacy
from spacy.matcher import PhraseMatcher
nlp = spacy.load("en_core_web_md")
matcher = PhraseMatcher(nlp.vocab)
terms = ["Angela Merkel", "Donald Trump", "Alexis Tsipras"]
patterns = [nlp.make_doc(term) for term in terms]
matcher.add("politiciansList", None, *patterns)
doc = nlp("3 EU leaders met in Berlin. German chancellor Angela
Merkel first welcomed the US president Donald Trump. The following
day Alexis Tsipras joined them in Brandenburg.")
matches = matcher(doc)
for mid, start, end in matches:
print(start, end, doc[start:end])
```

5 заданий, проверяющих ИПК-1.1 Обнаруживает знания об актуальных направлениях междисциплинарных лингвистических исследований в избранной научной сфере.

Задание 1

1. Что из перечисленного ниже является примером задачи генерации текста в НЛП?
   а) Анализ настроения
   b) Распознавание именованных сущностей
   в) Перевод языка
   г) Резюме текста

2. Как называется процесс определения грамматической структуры предложения?
   а) Парсинг
   б) Токенизация
   в) Лемматизация
   г) Стемминг

3. Что из нижеперечисленного является примером языковой модели, используемой в НЛП?
а) BERT (двунаправленные кодирующие представления из трансформаторов)
б) Классификатор Наивного Байеса
в) Случайный лес
г) K-nearest neighbors (KNN)

4. Какова цель распознавания именованных сущностей в НЛП?
а) Извлечение и классификация именованных сущностей, таких как имена, местоположения и организации.
б) Классифицировать документы по заранее определенным категориям
с) Анализ настроения, выраженного в тексте
г) Генерировать связные и осмысленные предложения

5. Как называется процесс определения настроения, выраженного в тексте?
а) Анализ настроения
б) Классификация текстов
в) Топологическое моделирование
г) распознавание именованных сущностей

1. в
2. а
3. а
4. а
5. а

Задание 2

1. Что из перечисленного ниже является примером платформы для чатботов?
а) Facebook Messenger
б) Google Maps
с) Microsoft Excel
г) Adobe Photoshop

2. Какова цель управления контекстом в чат-ботах?
а) Для поддержания последовательного течения разговора
б) Анализ настроения пользователей
в) Генерировать креативные ответы
г) Для распознавания речи

3. Что из перечисленного ниже является ограничением чат-ботов?
а) Неспособность понимать человеческий язык
б) Ограниченный словарный запас и база знаний
в) Отсутствие визуальных возможностей
г) Сложность одновременной работы с несколькими пользователями

4. Как называется процесс передачи разговора от чатбота к человеческому агенту?
а) Передача
б) Переход
в) переключение
г) передача

5. В чем заключается одно из ключевых преимуществ использования чат-ботов в обслуживании клиентов?
   а) Доступность 24/7
   б) Человекоподобные эмоции
   в) Физическая помощь
   г) автономная функциональность

1. а
2. а
3. б
4. а
5. а

Задание 3
Векторизируйте произвольный текст с помощью библиотеки spacy и выведите вектор.

Ответ
```
import spacy
nlp = spacy.load("en_core_web_md")
doc = nlp("I ate a banana.")
doc[3].vector
```

Задание 4
Напишите код, использующий предобученные векторные модели spacy, для определения семантического сходства нескольких слов.

Ответ
```
sentences = nlp("I purchased a science fiction book last week. I
loved everything related to this fragrance: light, floral and
feminine... I purchased a bottle of wine. ")
key = nlp("perfume")
for sent in sentences.sents:
print(sent.similarity(key))
```

Задание 5
Напишите код, чтобы посчитать количество сущностей в корпусе.

Ответ
```
from collections import Counter
import spacy
nlp  = spacy.load("en_core_web_md")

corpus = open("data/atis_utterances.txt", "r").read().split("\n")

all_ent_labels = []
for sentence in corpus:
    doc = nlp(sentence.strip())
    ents = doc.ents
    all_ent_labels += [ent.label_ for ent in ents]

c = Counter(all_ent_labels)
print(c)
```

5 заданий, проверяющих 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.

Задача 1.
Напишите код, использующий библиотеку Whisper, транскрибирующий запись в текст.

Ответ
pip install openai-whisper

```
import whisper
model = whisper.load_model("base")
def transcribe(file):
print(file)
transcription = model.transcribe(file)
return transcription["text"]
```

Задача 2.
Напишите промпты для определения интента подаваемой ChatGPT последовательности

Ответ
```
prompts = {
"START": "Classify the intent of the next input. \
Is it: WRITE_EMAIL, QUESTION, OTHER ? Only answer one word.",
"QUESTION": "If you can answer the question: ANSWER, \
if you need more information: MORE, \
if you cannot answer: OTHER. Only answer one word.",
"ANSWER": "Now answer the question",
"MORE": "Now ask for more information",
"OTHER": "Now tell me you cannot answer the question or do the action",
"WRITE_EMAIL": 'If the subject or recipient or message is missing, \
answer "MORE". Else if you have all the information, \
answer "ACTION_WRITE_EMAIL |\
subject:subject, recipient:recipient, message:message".',
}
```

Задача 3.
Напишите промпт, согласно которому ChatGPT проверит текст на соответствие норме.

Ответ
Correct this to standard English: She no went to the market.

Задача 4.
Напишите промпт, согласно которому ChatGPT объяснит, что делает программный код.

Ответ
```
# Python 3
def hello(x):
print('hello '+str(x))
# Explanation of what the code does
```

Задача 5
Дайте описание роли нутрициолога для ChatGPT и задание для составления меню на обед.

Ответ
```
prompt = """
Role: You are a nutritionist designing healthy diets for high-performance
athletes. You take into account the nutrition needed for a good recovery.
Context: I do 2 hours of sport a day. I am vegetarian, and I don't like green
vegetables. I am conscientious about eating healthily.
Task: Based on your expertise defined in your role, give me a suggestion for a
main course for today's lunch. With this suggestion, I also want a table with
two columns where each row in the table contains an ingredient from the main
course. The first column in the table is the name of the ingredient.
The second column of the table is the number of grams of that ingredient needed
for one person. Do not give the recipe for preparing the main course.
"""
chat_completion(prompt)
```

5 заданий, проверяющих ИПК-4.2 Разрабатывает программу действий по решению задач проекта в области когнитивной и компьютерной лингвистики с учетом имеющихся технических средств и информационных технологий, в том числе в области искусственного интеллекта.
Задача 1
Напишите программный код, позволяющий посмотреть произвольное количество датасетов Huggingface.

Ответ
```
from datasets import list_datasets

all_datasets = list_datasets()
print(f"There are {len(all_datasets)} datasets currently available on the Hub")
print(f"The first 10 are: {all_datasets[:10]}")
There are 1753 datasets currently available on the Hub
The first 10 are: ['acronym_identification', 'ade_corpus_v2', 'adversarial_qa',
'aeslc', 'afrikaans_ner_corpus', 'ag_news', 'ai2_arc', 'air_dialogue',
'ajgt_twitter_ar', 'allegro_reviews']
```

Задача 2
Напишите пример кода, использующего фреймворк Langchain для более удобной работы с промптами

Ответ
```
from langchain.chat_models import ChatOpenAI
from langchain import PromptTemplate, LLMChain

template = """Question: {question}
Answer: Let's think step by step."""
prompt = PromptTemplate(template=template, input_variables=["question"])

llm = ChatOpenAI(model_name="gpt-3.5-turbo")
llm_chain = LLMChain(prompt=prompt, llm=llm)
```

question = "What is the population of the capital of the country where the Olympic Games were held in 2016?"
llm_chain.run(question)

Задача 3
Напишите пример кода, где Langchain интегрируется с OpenAI API для генерации текста

Ответ
```
from langchain.chat_models import ChatOpenAI
from langchain import PromptTemplate, LLMChain
template = """Question: {question}
Let's think step by step.
Answer: """
prompt = PromptTemplate(template=template, input_variables=["question"])
llm = ChatOpenAI(model_name="gpt-4")
llm_chain = LLMChain(prompt=prompt, llm=llm)
question = """ What is the population of the capital of the country where the
Olympic Games were held in 2016? """
llm_chain.run(question)
```

Задача 4
Охарактеризуйте взаимоотношения между обучающим, тестовым и валидизирующим набором данных, используя блок-схему.

Ответ

Задача 5
Напишите такой код, который удивит Вашего преподавателя!

Ответ
```
class Solution:
    def minMalwareSpread(self, graph: List[List[int]], initial: List[int]) -> int:
        n=len(graph)
        par=[-1]*n
        size=[1]*n

        for i in range(n):
            par[i]=i
```

```
def findpar(i):
    if par[i]==i:
        return i
    else:
        par[i]=findpar(par[i])
        return par[i]

def merge(i1,i2):
    if size[i1]>=size[i2]:
        par[i2]=i1
        size[i1]+=size[i2]
    else:
        par[i1]=i2
        size[i2]+=size[i1]

for i in range(n):
    for j in range(n):
        if graph[i][j]==1:
            p1=findpar(i)
            p2=findpar(j)
            # print(i,p1,j,p2)
            if p1!=p2:
                merge(p1,p2)
                # print(par[i],size[i],par[j],size[j],56)

inf=[0]*n
for i in initial:
    p=findpar(i)
    inf[p]+=1

mini=-1
ans=-1
# print(par)
# print(size)
initial.sort()
for i in initial:
    # print(i)
    p=findpar(i)
    # print(i,p,size[p])
    if inf[p]==1 and size[p]>mini:
        ans=i
        mini=size[p]
            return initial[0] if ans==-1 else ans
```

Информация о разработчиках

Шиляев Константин Сергеевич, к. филол. н., доцент, кафедра общей, компьютерной и когнитивной лингвистики