

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

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Оценочные материалы по дисциплине

Интеллектуальный анализ текста

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки:
Обработка данных, управление и исследование сложных систем

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
Л.А. Нежелская

Председатель УМК
С.П. Сущенко

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

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.1 Анализирует проблемы в области фундаментальной и прикладной математики.

ИОПК-1.3 Решает актуальные задачи фундаментальной и прикладной математики.

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

- практические работы;
- зачет.

Задания для практических работ (ИОПК-1.1, ИОПК-1.3)

1. Практическая работа «Предобработка текста»

Требуется прочитать текст на русском языке из файла и вывести все пары соседних слов, которые:

- имеют имена существительные или имена прилагательные на первом или втором месте;
- совпадают по роду, числу и падежу.

Все пары следует выводить в виде лемм. Например, если исходная пара имела вид «необычайных университетов», то должна быть выведена пара «необычайный университет».

2. Практическая работа «Векторное представление слов»

Используя `import gensim`, необходимо реализовать вычисление десяти самых близких по смыслу слов, находящихся в окрестности от результата операций сложения и вычитания в векторной модели. Каждому студенту преподавателем будет дана пара слов и необходимо найти такую линейную комбинацию исходных слов, чтобы в результате вычислений заданная пара попадала в первую десятку.

3. Практическая работа «Анализ на основе RNN»

Необходимо провести повторный анализ текста, который использовался в работе «Предобработка текста», но с использованием `import rnnmorph` и без использования `import rumorphu2`. Если полученные результаты различаются, необходимо пояснить, почему так вышло.

4. Практическая работа «Маскирование слов»

Используя модель BERT и её функцию `Masked language modelling`, требуется реализовать вычисление десяти самых вероятных слов, на месте любого умышленно пропущенного слова в корректно составленном предложении на русском языке.

Каждому студенту преподавателем будет дана пара слов, и требуется построить окружение, т. е. само возможное предложение на русском языке с пропущенным словом, для которого в вариантах подстановки пара слов будет встречаться в первой десятке. Слова должны совпадать с точностью до словоформы (слово «домами» не может подходить под требуемое слово «домом»).

5 Практическая работа «Генерация текста»

Используя модель RuGPT от Сбера, необходимо реализовать возможность генерации текста по заданному промпту. Допускается использование как старой модели `rugpt3large_based_on_gpt2`, так и новой `ruGPT-3.5-13B`.

Каждому студенту преподавателем будет дана пара слов, и требуется подобрать промпт таким образом, чтобы выданная пара слов встречалась бы в сгенерированном тексте с учётом порядка и с учётом словоформ (как в предыдущей работе). Допускается использовать обе модели, но пара слов преподавателем подбирается на `rugpt3large_based_on_gpt2`. Для любой модели допускается использовать режим `Diverse beam search decoding`. Ограничений на значения параметров нет (даже на длину генерируемого текста).

Функция `generate` для практической работы «Генерация текста»:

```
def generate(
    model, tok, text,
    do_sample=True, max_length=100, repetition_penalty=5.0,
    top_k=5, top_p=0.95, temperature=1,
    num_beams=None,
    no_repeat_ngram_size=3
):
    input_ids = tok.encode(text, return_tensors="pt")
    print(model.generate.__globals__['__file__'])
    out = model.generate(
        input_ids,
        max_length=max_length,
        repetition_penalty=repetition_penalty,
        do_sample=do_sample,
        top_k=top_k, top_p=top_p, temperature=temperature,
        num_beams=num_beams, no_repeat_ngram_size=no_repeat_ngram_size
    )
    return list(map(tok.decode, out))
```

Критерии оценивания:

Зачёт за практическую работу ставится, если даны верные ответы, и автор способен пояснить полученный результат.

3. Оценочные материалы итогового контроля и критерии оценивания

На экзамене будет сгенерирован билет, в который войдут три случайные темы из списка ниже (ИОПК-1.1):

1. Типы задач обработки естественного языка; распространённые варианты предобработки текста; трудности в их реализации.
2. Формальные аналитические грамматики; вероятностные модели; СММ; алгоритм Витерби.
3. Перцептрон; линейная ячейка; уравнение линейной ячейки; функция активации; порядок обучения нейронной сети.
4. Векторное представление слов; алгебраические операции над словами; модель `Word2vec`.
5. Рекуррентная сеть; модель Элмана; уравнения модели Элмана для последовательностей и для вектора скрытого состояния; недостатки рекуррентных сетей.

6. Модель Seq2seq; кодер и декодер; задачи для Seq2seq; долгая краткосрочная память.
7. Идея механизма внимания; внимание Богданова и Луонга; внутреннее внимание; multi-head attention/.
8. Трансформер; позиционное кодирование; преимущества Трансформера; взаимосвязь кодера и декодера Трансформера.
9. BERT, GPT и прикладные вопросы использования генеративных моделей.
10. Идея генерации изображений и речи; диффузионная модель; MFCC; Tacotron2.

Критерии оценивания:

Зачет выставляется, если сданы практические работы, на все теоретические вопросы даны правильные развернутые ответы.

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

Теоретические вопросы (ИОПК-1.1)

1. Типы задач обработки естественного языка; распространённые варианты предобработки текста; трудности в их реализации.
2. Формальные аналитические грамматики; вероятностные модели; СММ; алгоритм Витерби.
3. Векторное представление слов; алгебраические операции над словами; модель Word2vec.
4. Модель Seq2seq; кодер и декодер; задачи для Seq2seq; долгая краткосрочная память.
5. Идея механизма внимания; внимание Богданова и Луонга; внутреннее внимание; multi-head attention/.
6. BERT, GPT и прикладные вопросы использования генеративных моделей.
7. Идея генерации изображений и речи; диффузионная модель; MFCC; Tacotron2.

Информация о разработчиках

Пожидаев Михаил Сергеевич, канд. техн. наук, доцент кафедры теоретических основ информатики.