Инструкция по защите проектов

Итоговая аттестация

Цель: Разработка и тестирование приложения, основанного на машинном обучении.

Задачи:

- Анализ существующих решений для выбранной темы
- Запуск базовых моделей
- Оценка качества результата по релевантным для задачи метрикам
- Получение отчетов по результатам

Форма работы: индивидуальная / групповая (2 человека в группе)

Набор технологий:

- Python, Pandas, Numpy
- Pytorch / Tensorflow / Keras
- OpenCV, NLTK, SpaCy, Natasha, платформа nvidia jetson nano
- Google Colaboratory, Docker, Flask / Django
- Другие подходящие для задачи библиотеки

План работы:

- Необходимо провести предварительный анализ существующих решений, доступных наборов данных, включая их сравнение.
- Необходимо разработать базовую программную реализацию модели (допускается использование существующей реализации)
- Необходимо провести эксперименты с одним или более наборами данных и представить результаты в виде отчета в формате PDF

Проекты на выбор:

Проект 1. Система рекомендаций по фильмам, основанная на исходящих ссылках из Википедии.

В типичной системе рекомендаций мы даем рекомендации, основанные на нескольких фильмах, которые оценил пользователь.

Необходимо: изучить обучающий набор данных из Википедии, обучить эмбеддинги для фильмов на основе ссылок между статьями. Это можно сделать, обучив сеть, которая предсказывает фильм на основе исходящих ссылок на соответствующей странице Википедии. Затем реализовать нужно классификатор (например, SVM), чтобы давать рекомендации о фильмах (использовать расстояние от разделяющей гиперплоскости как меру полезности для пользователя).

Данные: wp_movies_10k.ndjson

https://drive.google.com/drive/folders/1r9KZPqUBcuDnyvNoJCFfEd4RJY1Z2jXL?usp=sharing

Проект 2. Система, предлагающая смайлики на основе фрагмента текста.

Простой проект на основе имеющегося датасета: твиты + смайлики

Необходимо: Разработать классификатор тональности, основанный на общедоступном наборе твитов, помеченных различными эмоциями, такими как счастье, любовь, удивление и т.д. Затем натренировать сверточную сеть и рассмотреть различные способы настройки этого классификатора. На вход модели приходит текст твита, на выходе: эмодзи.

Данные: emojis.csv

https://drive.google.com/drive/folders/1r9KZPqUBcuDnyvNoJCFfEd4RJY1Z2jXL?usp=sharing

Проект 3. Решить задачу DaNetQA / BoolQ

DaNetQA - это набор да/нет вопросов с ответами и фрагментом текста, содержащим ответ. Все вопросы были написаны авторами без каких-либо искусственных ограничений. Каждый пример представляет собой триплет (вопрос, фрагмент текста, ответ) с заголовком страницы в качестве необязательного дополнительного контекста. Настройка классификации текстовых пар аналогична существующим задачам логического вывода (NLI). Можно решить как задачу для русского, так и для английского. Либо провести эксперименты с многоязычной моделью.

Датасет и описание:

https://russiansuperglue.com/ru/tasks/task_info/DaNetQA

Проект 4. Решить задачу извлечения именованных сущностей для русского

Необходимо: обучить и протестировать модель для извлечения именованных сущностей из текста. Провести анализ решения и альтернатив. Выбрать лучшую модель.

Датасеты:

http://bsnlp.cs.helsinki.fi/shared-task.html

https://multiconer.github.io

Проект 5. Поиск похожих картинок (цветов)

Необходимо: обучить и протестировать модель для поиска похожих картинок. Коллекции для поиска и обучения нужно собрать из предложенных ниже наборов данных.

Датасеты:

https://www.kaggle.com/alxmamaev/flowers-recognition

https://www.kaggle.com/c/tpu-getting-started/data

https://www.robots.ox.ac.uk/vgg/data/flowers/102/index.html

Проект 6. Генератор анекдотов

Необходимо: разработать модель генерации анекдотов и (дополнительно) интерфейс для ее использования (например, бот Telegram). Готовая система генерирует случайные анекдоты по запросу или берет начало анекдота и завершает его. Используйте модель GPT-2.

Датасет:

anecdotes.csv

https://drive.google.com/drive/folders/1r9KZPqUBcuDnyvNoJCFfEd4RJY1Z2jXL?usp=sharing

Проект 7. Вопросно-ответный поиск

Необходимо: обучить и протестировать модель для поиска ответа на вопросы. На входе вопрос пользователя, система ищет похожий вопрос в базе вопросов с ответами и выдает ответ. Провести анализ решения и альтернатив. Выбрать лучшую модель.

Датасет: TBA (для английского можно использовать базу stackexchange)

Критерии проекта:

Код должен быть выложен на github / Google Colaboratory и удовлетворять следующим критериям:

- Оценка за код задания будет распределена между следующими аспектами:
    - функциональность,
    - структура и организация кода,
    - инструкция для запуска моделей.

Оценка отчета и презентации состоит из следующих компонент:

- качество отчета,
- качество документации по наборам данных,
- качество слайдов с постановкой задачи, выбранным подходом и результатами

Структура отчета[1]:

- Часть 1. Введение
- Часть 2: Обзор литературы
- Часть 3: Методология: включая план экспериментов, применяемые статистические методы.
- Часть 4: Результаты применения моделей и методов

Шкала оценивания (зачет/незачет):

Оценки 1/«отлично» заслуживает работа, в которой полностью и всесторонне раскрыто содержание программы обучения, обоснован выбор модели, представлен работающий код, содержится творческий подход к решению вопросов, сделаны обоснованные предложения и на все вопросы

[1] Нет смысла добиваться толстых отчетов и большого количества страниц. Будьте лаконичны и пишите по делу

при защите слушатель дал аргументированные ответы. Проект соответствует указанным показателям.

Оценки 0.8/«хорошо» заслуживает работа, в которой содержание изложено на высоком уровне, правильно сформулированы выводы и даны обоснованные предложения, на все вопросы слушатель дал правильные ответы. Проект в большей степени соответствует указанным показателям.

Оценки 0.5/«удовлетворительно» заслуживает работа, в которой в основном раскрыто содержание программы обучения, выводы в основном правильные. Предложения представляют интерес, но недостаточно аргументированы и на все вопросы слушатель дал правильные ответы. Проект в целом соответствует указанным показателям.

Оценки 0/«неудовлетворительно» заслуживает работа, которая в основном раскрывает поставленную тему, но при защите слушатель не дал правильных ответов на большинство заданных вопросов, то есть обнаружил серьезные пробелы в профессиональных знаниях, либо в проекте не проведено ни одного эксперимента. Проект не соответствует указанным показателям.