Инструкция по защите проектов

Итоговая аттестация

Цель: Разработка и тестирование приложения для промышленности, 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

Задачи и отчетность

Для успешного завершения вам требуется подготовить решение в виде

- Jupyter notebook или Flask/Django приложение. Программная реализация должна включать рабочий, выполняемый код и комментарии, графики и пр.

- Отчет в формате PDF, который должен включать Введение, Постановку задачи, Обзор моделей, Описание процесса решения и Результаты. Объем (от 2 до 10 стр.)

Темы проектов на выбор:

1. Разработка и тестирование приложения для автоматического анализа данных в промышленности с помощью нейронных сетей.

2. Создание и тестирование приложения для классификации изображений в промышленности с помощью машинного обучения.

3. Разработка и тестирование приложения для анализа текста в промышленности с использованием машинного обучения.

4. Разработка и тестирование приложения для анализа и прогнозирования данных в промышленности с помощью нейронных сетей.

5. Разработка и тестирование приложения для создания и обучения моделей глубокого обучения в промышленности.

6. Разработка и тестирование приложения для прогнозирования будущих событий с использованием машинного обучения в промышленности.

7. Разработка и тестирование приложения для анализа больших данных в промышленности с использованием машинного обучения.

8. Разработка и тестирование приложения для распознавания голоса в промышленности с помощью машинного обучения.

9. Разработка и тестирование приложения для создания автоматических диагностических систем в промышленности с использованием машинного обучения.

10. Разработка и тестирование приложения для мониторинга производства и предсказания времени производства с использованием машинного обучения.

11. Разработка и тестирование приложения для распознавания образов в промышленности с использованием машинного обучения.

12. Разработка и тестирование приложения для анализа и прогнозирования транспортных данных с использованием машинного обучения.

13. Разработка и тестирование приложения для автоматического создания программ в промышленности с использованием машинного обучения.

14. Разработка и тестирование приложения для распознавания речи в промышленности с использованием машинного обучения.

15. Разработка и тестирование приложения для автоматического поиска и анализа данных в промышленности с помощью машинного обучения.

16. Разработка и тестирование приложения для распознавания движений на промышленных объектах с использованием машинного обучения.

17. Разработка и тестирование приложения для распознавания рукописных данных в промышленности с использованием машинного обучения.

18. Разработка и тестирование приложения для анализа и прогнозирования метеорологических данных, влияющих на промышленные объекты, с использованием машинного обучения.

19. Разработка и тестирование приложения для автоматического поиска закономерностей в промышленности с использованием машинного обучения.

20. Разработка и тестирование приложения для автоматического анализа социальных данных с помощью машинного обучения.

21. Использование машинного обучения для оптимизации расходов и ресурсов.

22. Разработка безопасного машинного обучения для промышленных приложений.

23. Тестирование машинного обучения для промышленных приложений.

24. Использование машинного обучения для улучшения производительности промышленных приложений.

25. Использование машинного обучения для оптимизации рабочих процессов.

Примеры реализации проектов:

Проект 1. Система рекомендаций по подбору промышленных узлов и агрегатов, основанная на исходящих ссылках из Википедии.

В типичной системе рекомендаций мы даем рекомендации по подбору промышленных узлов и агрегатов, основанные на нескольких сборных моделях, которые описал пользователь.

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

Данные: wp_movies_10k.ndjson (пример реализации по статьям) https://github.com/DOsinga/deep_learning_cookbook/blob/master/data/wp_movies_10k.ndjson

Проект 2. Система, предлагающая маркировку на промышленные узлы и агрегаты на основе фрагмента текста.

Простой проект на основе имеющегося датасета: твиты + графические изображения

Необходимо: Разработать классификатор тональности, основанный на общедоступном наборе твитов, помеченных различными графическими изображениями. Затем натренировать сверточную сеть и рассмотреть различные способы настройки этого классификатора. На вход модели приходит текст твита, на выходе: графическое изображение для маркировки промышленных узлов и агрегатов.

Данные: img_gafics.csv

https://gist.github.com/bfeldman89/fb25ddb63bdaa6de6ab7ac946acde96f

Альтернативная ссылка:

https://figshare.com/articles/dataset/smile_annotations_final_csv/3187909

второй датасет (посложнее) :

https://uvaauas.figshare.com/articles/dataset/Twemoji_Dataset/5822100

Проект 3. Решить задачу DaNetQA / BoolQ в промышленной отрасли

DaNetQA - это набор да/нет вопросов с ответами и фрагментом текста, содержащим ответ в промышленной отрасли. 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 в промышленной отрасли

Необходимо: обучить и протестировать модель для извлечения именованных сущностей из текста в промышленной отрасли. 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) в промышленной отрасли

Необходимо: обучить и протестировать модель для поиска похожих картинок в промышленной отрасли. 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 аналитических справок по промышленным объектам

Необходимо: разработать модель генерации аналитических справок по промышленным объектам и (дополнительно) интерфейс для ее использования (например, бот Telegram). Готовая система генерирует случайные справки по запросу или берет описания объекта и завершает его. Используйте модель GPT-2

Датасет:
promobjects.csv
https://www.kaggle.com/datasets/konstantinalbul/promobjects

Проект 7. Вопросно-ответный поиск в промышленной отрасли
Необходимо: обучить и протестировать модель для поиска ответа на вопросы в промышленной отрасли. 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 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

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

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.