Санкт-Петербургский государственный университет

Кафедра информационно-аналитических систем

Валидатор текстов выпускных квалификационных работ

Автор: Макеев Владислав Дмитриевич, 21.М05-мм

Научный руководитель: Чернышев Георгий Алексеевич

Санкт-Петербургский государственный университет

27 ноября 2021г.

Оглавление

1. Введение 3
2. Используемые технологии 4

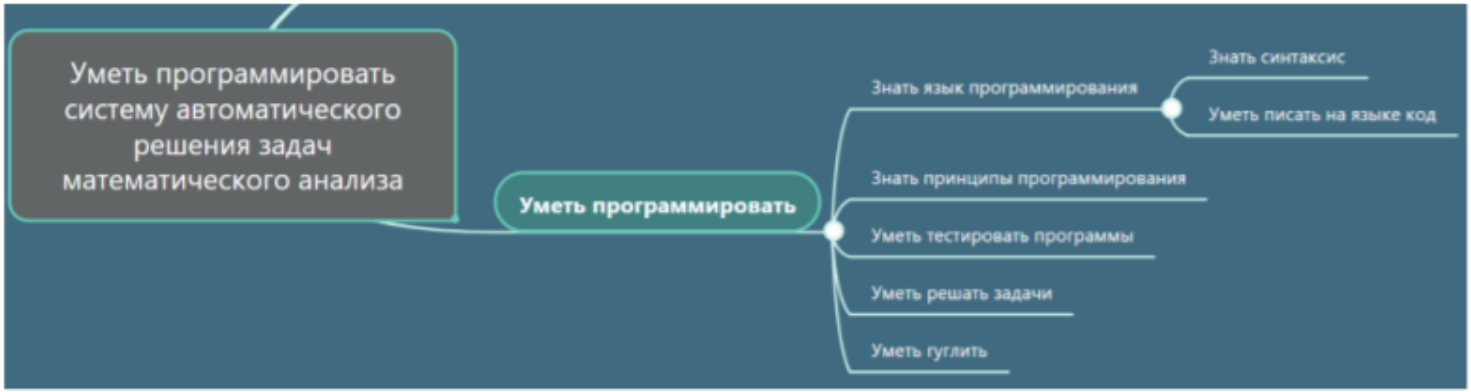


Рисунок 3 – пример дерева компетенций

- SrcSeq – стандартная модель

3. Выбор датасета **17**
3.1. 150k Python Dataset 17

4. Предобработка данных **19**

Введение

Оглавление

1. Введение 3
2. Используемые технологии 4

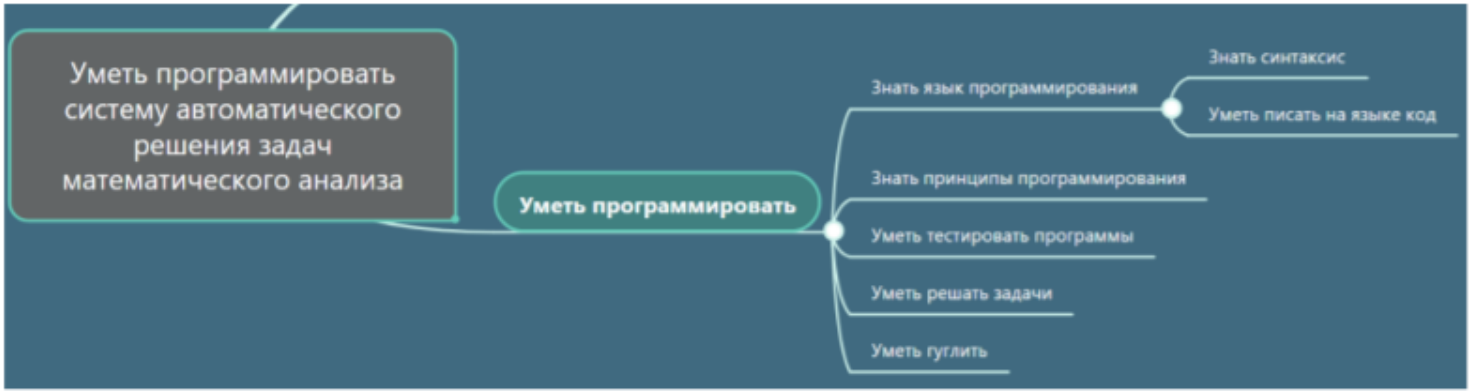


Рисунок 3 – пример дерева компетенций

- SrcSeq – стандартная модель

3. Выбор датасета **17**
3.1. 150k Python Dataset 17

4. Предобработка данных **19**

Введение

Оглавление

1. Введение 3
2. Используемые технологии 4

Рисунок 3 – пример дерева компетенций

- SrcSeq – стандартная модель

3. Выбор датасета **17**
3.1. 150k Python Dataset 17

4. Предобработка данных **19**

Введение

Оглавление

1. Введение 3
2. Используемые технологии 4

Рисунок 3 – пример дерева компетенций

- SrcSeq – стандартная модель

3. Выбор датасета **17**
3.1. 150k Python Dataset 17

4. Предобработка данных **19**

Введение

Оглавление

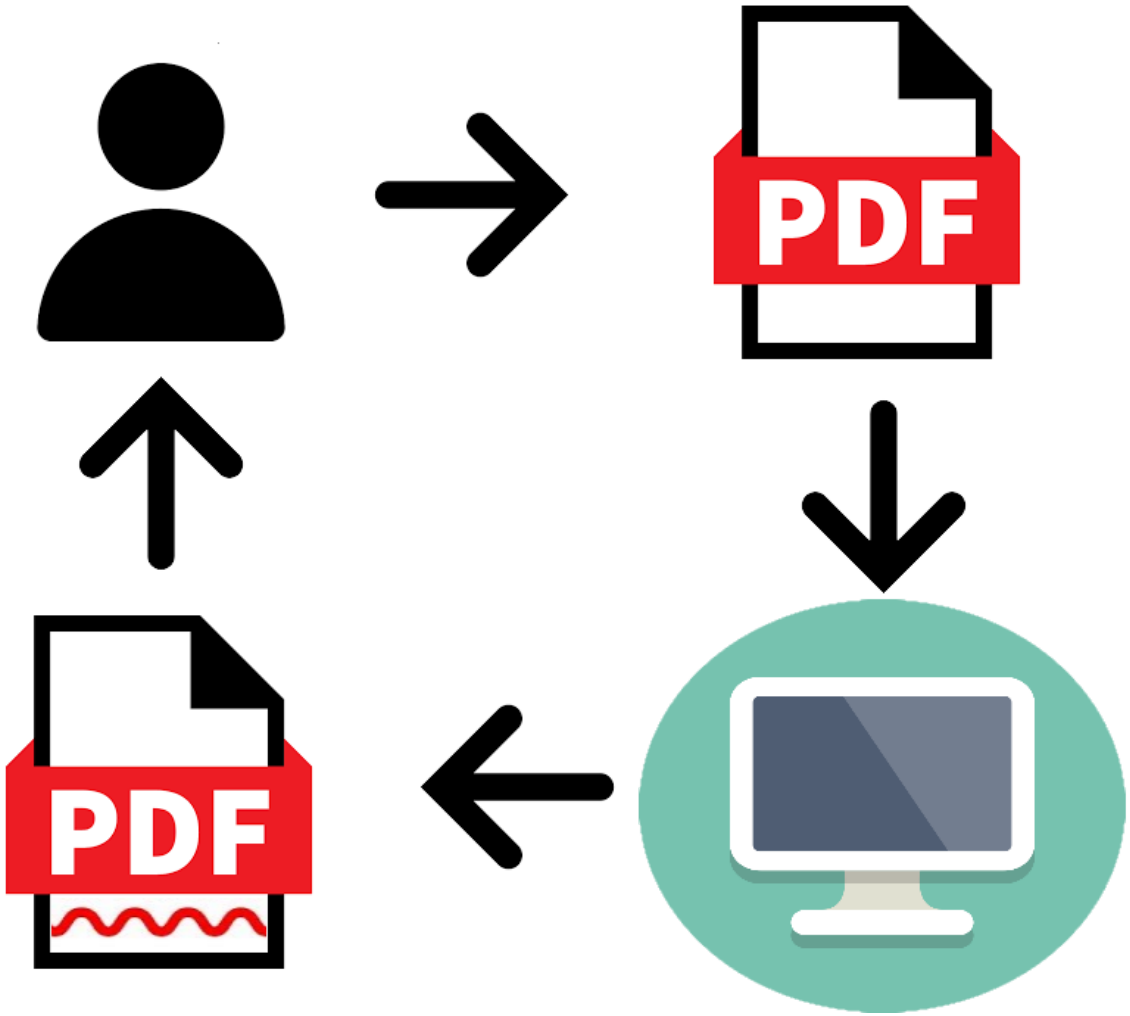
1. Введение 3
2. Используемые технологии 4

Рисунок 3 – пример дерева компетенций

- SrcSeq – стандартная модель

3. Выбор датасета **17**
3.1. 150k Python Dataset 17
4. Предобработка данных **19**

Введение



Что есть в проекте

Целью проекта является разработка веб-приложения, способного проверять соответствие PDF-файлов квалификационных работ определённому набору требований.

На данный момент:

- Разработан прототип веб-приложения, позволяющий:
  - осуществлять загрузку и просмотр PDF;
  - находить и оповещать пользователя о некоторых ошибках:
    - некорректное использование -, – и —;
    - “сломанные” ссылки на литературу [?];
    - списки из одного элемента;
  - осуществлять добавление аннотаций (подчёркивание ошибок);
  - анализировать структуру PDF (наличие оглавления, списка литературы).
- Автоматизирована сборка и тестирование

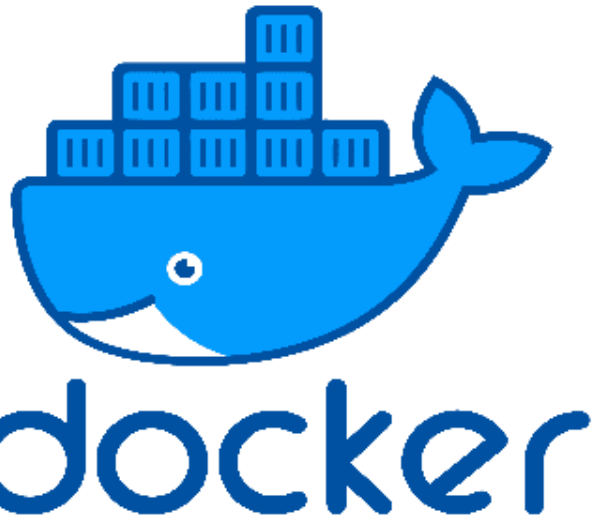
Дальнейшие планы

- Добавление дополнительных правил
- Добавление правил с помощью высокоуровневых средств
- Подключение сторонних библиотек машинного обучения
- Подключение сторонних библиотек для обработки естественных языков

Дополнительные правила

- Центрирование рисунков и подписей
- Корректные ссылки на рисунки
- Предложения в скобках с маленькой буквы
- Формат ссылок на литературу
- Отсутствие написанного вручную оглавления
- Корректные названия секций (капитализация, отсутствие специальных символов)
- Отсутствие формул и кода в виде рисунков

Инструментарий



Github

https://github.com/Darderion/map

Usage

- Docker:
  - `$ docker build -t map .`
  - `$ docker run --rm -it -p 8080:8080 map:latest`
- Maven:
  - `$ bash buildscript.sh`
  - `$ cd target`
  - `$ java -jar app.jar`

Languages

Kotlin 68.8% Vue 23.2%
TypeScript 5.4% HTML 0.9%
Shell 0.8% Dockerfile 0.8%
JavaScript 0.1%

Сценарий демонстрации

1. Пример для демонстрации работоспособности
2. Примеры реальных работ:
   - Работа с некорректным использованием дефиса
   - Работа с некорректным использованием короткого тире
   - Работа с одним подразделом в разделе