УСТАВ ПРОЕКТА

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Наименование проекта** | HACKATON.UNIGOTO | | **Планируемое время начала и окончания проекта (месяц/год)** | 27.11.2023 – 15.12.2023 | | **Оценка бюджета проекта (млн/тыс. руб.)** | 0 | | **Место/сфера реализации** | IT->Рекомендательные системы и алгоритмы/  Образование -> сервисы помощи абитуриентам | | **Заказчик проекта** | UNIGOTO | | **Другие/ключевые участники проекта** | Афлатунов Эмиль, Шухардин Александр, Гайфуллин Александр | | **Дата создания документа** | 06.12.2023 | |

**Цели проекта**

Итоговая цель проекта: реализовать рекомендательную систему на основе предоставленных заказчиком данных, способную давать релевантные рекомендации пользователям на основе получаемой от них обратной связи по определенным личным параметрам (увлечения, интересы любимые книги и т. д.)

**Описание решения.**

Основная концепция подхода к решению задачи алгоритмы заключается в применении алгортма векторизации TF-IDF (Term Frequency – Inverse Document Frequence), а также фреймворка SentenceTransformers, в качестве модели для которого обучается и используется для рекомендаций 'DeepPavlov/rubert-base-cased'.

**1й этап:** формирования исходного датасета с помощью API, предоставленного заказчиками проекта

**2й этап:** проведено несколько этапов обработки данных: формирование репрезентативной выборки, очищенной от пропусков данных и нерелевантной для задачи информации с помощью методов EDA.

**3й этап:** Имитация пользовательского ввода данных с помощью методов Google Colab

**4й этап:** Реализация алгоритма рекомендательной системы на основе подхода TF-IDF векторизации с помощью вычисления косинусной близости между векторизованными данными пользовательского ввода и очищенного датасета с использованием sklearn.metrics.pairwise, TfidfVectorizer, nltk, pymorphy2.MorphAnalyzer.

**5й этап:** Обучение модели 'DeepPavlov/rubert-base-cased', получение text-embeddings, вычисление косинусной близости между пользовательским вводом и вычисленным text-embeddings.

**6й этап:** Представление результатов.

**Применимость.**

Данный проект будет полезен абитуриентам, выбирающим вуз для получения высшего образования.

**Команда.**

Афлатунов Эмиль – разработка алгоритма решения задачи

Шухардин Александр – разработка алгоритма решения задачи

Гайфуллин Александр – оформление документации.

**Заключение.**

Реализованы алгоритмы построения рекомендательной системы на основе данных очищенного датасета.

|  |
| --- |
| **Требования к проекту и продукту**   1. Произвести первоначальную обработку данных заказчика, получить датасет, очищенный от неинтерпретируемых данных. 2. В перспективе должен быть реализован пользовательский ввод данных и вывод топ-списка из n (параметр также задается пользоавтелем) элементов с наиболее релевантными университетами и направлениями подготовки. 3. Предусмотрена возможность выдачи результатов по вузам, факультетам и специальностям по отдельности, а также с ограничением по городу. 4. Подробное описание всех шагов по обработке данных и настройке модели РС**.** |
| **UNIGOTO the goal and that's it!** |