Описание ПО программы BAUM AI PLATFORM

1. Назначение программы

Платформа прикладного ИИ BAUM AI представляет контейнер-ориентированную архитектуру, связывающую микросервисы, упакованные в docker контейнеры. Платформа поделена на модули, которые представляют из себя несколько контейнеров, объединенных в pod'ы. Оркестрация контейнерами выполняется с использованием специализированного ПО.

2. Описание программных модулей

Список программных модулей:

1. Модуль парсинга;
2. Модуль Data Quality;
3. Многомерная модель базы данных;
4. Модуль «базы данных»;
5. Модуль поиска на платформе;
6. Модуль предпроцессинга и разметки данных;
7. Модуль DA – data analysis (анализ данных);
8. Модуль ML – machine learning (машинное обучение);
9. Модуль DL – deep learning (глубокое обучение);
10. Модуль RL – reinforcement learning (обучение с подкреплением);
11. Модуль model saving & transfer;
12. Модуль конструктор AI drag n drop;
13. Модуль GUI – graphical user interface (графический интерфейс пользователя);
14. Модуль «Мастер-помощник».

001 Модуль парсинга

Модуль построен на базе Docker изолированной среды.

Модуль изолированно размещен в контейнере, с установкой всех дополнительных зависимостей и библиотек.

Модуль содержит набор функций и скриптов, необходимых для выполнения задач парсинга и краулинга данных из различных источников.

Модуль имеет возможность настройки работы с брокерами поставки данных - pache airflow, apache kafka.

Модуль размещен в одном «Pod 01» с модулем 002.

Модуль взаимодействует с модулями: 002, 003, 004 посредством kubernetes.

Рисунок 1. Pod 01

002 Модуль Data Quality

Модуль построен на базе Docker изолированной среды.

Модуль изолированно размещен в контейнере, с установкой всех дополнительных зависимостей и библиотек.

Модуль обеспечивает функцию качества данных (поступающих данных). Обеспечение качества данных осуществляется с использованием скриптов, и современных методов, включая машинное обучение.

Основные ошибки для исправления:

– Поиск опечаток и проверка орфографии

– Поиск выбросов и аномалий

– Сверка со справочниками (страны, почтовые индексы и т.д.)

– Проверка типов данных

Модуль размещен в одном «Pod 01» с модулем 001.

003 Многомерная модель базы данных

Многомерная модель базы отображает связи между основными компонентами (модулями) системы.

Модель базы данных выполнена в виде ER диаграммы, с указанием первичных (PK) и внешних ключей (FK), а также их связей.

Модель учитывает микросервисную архитектуру общей системы в целом.

004 Модуль «базы данных»

Модуль включает в себя основные базы данных, необходимые для работы с данными, в процессе машинного обучения и анализа.

Составные части модуля упакованы в единый «Pod 02», управляемый по средствам kubernetes.

Модуль состоит из отдельных независимых баз данных.

Предусмотрены:

База данных Postgresql – упакована в изолированный docker контейнер со всеми необходимыми зависимостями.

Обеспечивает хранение структурированных данных.

База данных Hbase – упакована в изолированный docker контейнер со всеми необходимыми зависимостями.

Обеспечивает поколоночное хранение неструктурированных данных.

База данных CuDF – упакована в изолированный docker контейнер со всеми необходимыми зависимостями.

База данных Rapida - упакована в изолированный docker контейнер со всеми необходимыми зависимостями.

База данных Apache Solr - упакована в изолированный docker контейнер со всеми необходимыми зависимостями.

База данных MongoDB - упакована в изолированный docker контейнер со всеми необходимыми зависимостями. Позволяет хранить программные и пользовательские скрипты, модели, блок-схемы.

Модуль хранения структурированных и неструктурированных данных

строится на базе ПО, с открытым исходным кодом, используя реляционные (postgreSQL), колоночные (habse), и/или документные (mongoDB) базы данных.

Модуль выполняет следующие функции:

а) хранение неструктурированных данных (тексты, изображения, звук);

б) хранение структурированных и параметризованных данных (числовые данные, временные ряды);

в) хранение пользовательских блок-схем, отчетов, моделей.

Модуль хранения структурированных и неструктурированных данных

обменивается данными с модулями:

а) интеграции с поисковой системой Directum;

б) drag n drop конструктора искусственного интеллекта и машинного обучения;

в) визуализации и формирования пользовательских отчетов;

г) модуль контроля версионности моделей (MLOps).

005 Модуль поиска на платформе

Модуль построен на базе поискового движка.

В качестве поисковых движков используются: Elastic Search и Apache Solr.

006 Модуль предпроцессинга и разметки данных

Модуль обработки и первичного анализа строится на базе существующего ПО, с открытым исходным кодом (pandas, numpy, sklearn, scipy), а также разработанных в рамках данного проекта библиотек, необходимых для выполнения прикладных задач обработки данных разных типов, структуры и размеров.

Модуль выполняется следующие основные функции:

а) анализ структуры, распределения и полноты данных

б) предварительная обработка, нормализация и стандартизация входных данных

в) преобразование, параметризация и подготовка данных для передачи в модели машинного обучения.

Модуль содержит методы для обработки и анализа различных типов входных данных:

– текстовые (предпроцессинг, очистка, лемматизация, кодирование, представление многомерными векторами)

– временные ряды (тесты на нормальность и стационарность ряда, декомпозиция ряда, приведение ряда к стационарному)

– числовые и категориальные (заполнение пропущенных данных, проверка на нормальность, нормализация и стандартизация, выявление аномалий, проверка на сбалансированность и ребалансировка, выделение наиболее важных признаков, кодирование, генерация новых признаков)

– графические (приведение к единой цветовой палитре, обрезка, дополнение данных и генерация новых образцов, бинарная сегментация).

Модуль библиотек и скриптов для обработки и предпроцессинга обменивается данными с модулями:

а) Модуль_007 анализ данных

б) Модуль_008 библиотек и алгоритмов машинного обучения

в) Модуль_009 библиотек нейронных сетей и глубокого обучения

г) Модуль_010 библиотек для обучения моделей «без учителя»

д) Модуль_013 графического интерфейса

е) Модуль_004 базы данных.

007 Модуль DA - data analysis

(в разработке)

008 Модуль ML - машинного обучения

Модуль библиотек и алгоритмов машинного обучения строится на базе ПО, с открытым исходным кодом (sklearn).

Модуль библиотек и алгоритмов машинного обучения выполняет следующие функции:

а) решение регрессионных задач

б) обучение вероятностных классификаторов

в) построение трендовых моделей

г) поиск аномалий

в) решение задач классификации (бинарной и множественной).

Модуль библиотек и алгоритмов машинного обучения обменивается данными с модулями:

а) хранения структурированных и неструктурированных данных

б) визуализации и формирования отчетов

в) графического интерфейса аналитического блока

г) «мастер-помощник»

д) drag n drop конструктор ИИ

е) контроля версионности (MLOps).

009 Модуль DL - deep learning

Модуль библиотек нейронных сетей и глубокого обучения строится на базе ПО, с открытым исходным кодом (tensorflow). Поддерживает Dense, Conv2D и LSTM слои нейронных сетей.

Модуль библиотек нейронных сетей и глубокого обучения выполняет следующие функции:

а) решение задач классификации (бинарная и множественная);

б) решение регрессионных задач;

в) определение вероятностей.

Модуль библиотек нейронных сетей и глубокого обучения обменивается данными с модулями:

а) хранения структурированных и неструктурированных данных;

б) визуализации и формирования отчетов;

в) графического интерфейса аналитического блока;

г) «мастер-помощник»;

д) drag n drop конструктор ИИ;

е) контроля версионности (MLOps).

010 Модуль RL reinforcement learning (обучение с подкреплением)

(в разработке)

011 Модуль model saving & transfer

Модуль контроля версионности моделей (MLOps) строится на базе ПО, с открытым исходным кодом (tensorflow-serving, tensorflow-extended).

Модуль контроля версионности моделей выполняет следующие функции:

а) сохранять модели и присваивать им версии;

б) обращаться к моделям соответствующей версии.

Модуль контроля версионности моделей обменивается данными с модулями:

а) модуль хранения структурированных и неструктурированных данных;

г) модуль библиотек и алгоритмов машинного обучения;

д) модуль библиотек нейронных сетей и глубокого обучения;

е) модуль библиотек исследований и обучения «без учителя»;

ж) модуль drag n drop конструктора искусственного интеллекта и машинного обучения.

012 Модуль конструктор AI drag n drop

Модуль drag n drop конструктора искусственного интеллекта и машинного обучения строится на базе написанного под проект ПО.

Модуль drag n drop конструктора искусственного интеллекта и машинного обучения выполняет следующие функции:

а) создание блок-схем машинного обучения и нейронных сетей, без прямого кодирования;

б) обеспечение взаимодействия между модулями, для решения аналитических задач;

в) сохранение блок-схем, для последующего использования и доработки.

Модуль drag n drop конструктора искусственного интеллекта и машинного обучения обменивается данными с модулями:

а) модуль интеграции с поисковой системой Directum;

б) модуль хранения структурированных и неструктурированных данных;

в) модуль библиотек и скриптов для обработки и предпроцессинга данных;

г) модуль библиотек и алгоритмов машинного обучения;

д) модуль библиотек нейронных сетей и глубокого обучения;

е) модуль библиотек исследований и обучения “без учителя”;

ж) модуль drag n drop конструктора искусственного интеллекта и машинного обучения;

з) модуль визуализации и формирования пользовательских отчетов;

и) модуль контроля версионности моделей (MLOps);

k) модуль графического интерфейса аналитического блока;

л) модуль «мастер-помощник».

013 Модуль GUI

Модуль визуализации и формирования пользовательских отчетов строится на базе ПО, с открытым исходным кодом.

Модуль визуализации и формирования пользовательских отчетов обеспечивает выполнение функций:

а) отображать ключевые показатели обученных моделей;

б) формировать отчеты о результатах обучения (график обучения, метрики ошибок и качества) модели и сохранять в формате pdf.

Модуль визуализации и формирования пользовательских отчетов обменивается данными с модулями:

а) модуль хранения структурированных и неструктурированных данных;

б) модуль drag n drop конструктора искусственного интеллекта и машинного обучения.

Модуль графического интерфейса аналитического блока строиться на базе ПО, написанного под проект на языке программирования.

Модуль графического интерфейса аналитического блока выполняет следующие функции:

а) обращение к соответствующим модулям в режиме GUI;

б) использование функционала модулей в режиме GUI.

Модуль графического интерфейса аналитического блока обменивается данными с модулями:

а) модуль интеграции с поисковой системой Directum;

б) модуль хранения структурированных и неструктурированных данных;

в) модуль библиотек и скриптов для обработки и предпроцессинга данных;

г) модуль библиотек и алгоритмов машинного обучения;

д) модуль библиотек нейронных сетей и глубокого обучения;

е) модуль библиотек исследований и обучения “без учителя”;

ж) модуль drag n drop конструктора искусственного интеллекта и машинного обучения;

з) модуль визуализации и формирования пользовательских отчетов;

и) модуль контроля версионности моделей (MLOps);

k) модуль графического интерфейса аналитического блока;

л) модуль «мастер-помощник».

014 Модуль «мастер-помощник»

Модуль «мастер-помощник» строится на базе ПО, написанного под проект на языке программирования.

Модуль «мастер-помощник» выполнять следующие функции:

а) формирование подсказок при создании блок-схем

б) формирование подсказок при обращении к базам данных

в) формирование подсказок при предпроцессинге данных

г) формирование подсказок при выборе методов и моделей для анализа данных

Модуль «мастер-помощник» обменивается данными с модулями:

а) модуль хранения структурированных и неструктурированных данных;

б) модуль библиотек и скриптов для обработки и предпроцессинга данных;

в) модуль библиотек и алгоритмов машинного обучения;

г) модуль библиотек нейронных сетей и глубокого обучения;

д) модуль библиотек исследований и обучения “без учителя”;

е) модуль drag n drop конструктора искусственного интеллекта и машинного обучения;

ж) модуль визуализации и формирования пользовательских отчетов;

з) модуль контроля версионности моделей (MLOps);

и) модуль графического интерфейса аналитического блока.