

Описание docker-image airpol

Данный docker образ содержит базовую систему для предсказаний индекса качества воздуха на 3 предыдущих дня для загрязнителя PM2.5 в городе Роттердам (Нидерланды).

Примечание. На настоящий момент выбран загрязнитель PM2.5, поскольку он является преобладающим большую часть года. В ближайшем будущем планируется доработать сервис таким образом, чтобы он возвращал предсказание на будущие 3 дня по двум загрязнителям (PM2.5 и O3), на основании моделей LightGBM, ARIMA и нейросети RNN.

1. Состав образа

1.1. Скрипты python для обновления данных и обучения модели

В систему входят следующие скрипты:

1.1.1. Скрипты для докачивания актуальных данных из внешних источников.

1.1.2. Скрипты для очистки, обогащения и фильтрации данных.

1.1.3. Скрипты для переобучения модели на актуализированных данных.

Данные скрипты выполняются автоматически при запуске контейнера, а также по таймеру ежедневно в 02:00.

1.2. Веб-сервис

Веб-сервис возвращает предсказания индекса качества воздуха. Он работает на внутреннем порту 8041.

Описание интерфейса можно получить после запуска контейнера по следующей ссылке:
[http://127.0.0.1:\[проброшенный при запуске порт\]/docs](http://127.0.0.1:[проброшенный при запуске порт]/docs)

По этому же адресу можно выполнить тестирование методов сервиса.

Сервис предоставляет следующие методы:

- | | |
|-----------|---|
| / | - Get.
Без параметров.
Формат ответа: строка Ok.
Используется для health-check сервиса |
| /predict | - Get.
Без параметров.
Формат ответа: массив из трех чисел с плавающей точкой.
Возвращает предсказание индекса качества воздуха по загрязнителю PM2.5, полученное на обученной после актуализации данных модели. |
| /evaluate | - Get.
Без параметров. |

Формат ответа: словарь, с двумя ключами `train_`[название используемой метрики] и `val_`[название используемой метрики] и двумя их значениями, представленными числами с плавающей точкой.

Возвращает метрики обучения и валидации модели, обученной на актуализированных данных.

1.3. Обработанные исходные данные за предыдущий период

В образ помещены данные за 2015-2022гг. После докачивания и обработки актуальных данных за 2023г они объединяются в один датасет, используемый для обучения модели. Это сделано для сокращения времени обработки данных.

2. Процедура запуска контейнера

1.1. Запуск контейнера

Образ размещен на [dockerhub](#). Имя образа: `alexandranasonova/airpol:1.0`

Для запуска контейнера можно воспользоваться следующим скриптом:

```
docker run -it -p [наружный порт]:8041 alexandranasonova/airpol:1.0
```

где в качестве наружно порта нужно указать любой свободный порт, на котором будет работать веб-сервис. Например так:

```
docker run -it -p 8041:8041 alexandranasonova/airpol:1.0
```

1.2. Процессы при запуске контейнера

При запуске контейнера будут автоматически выполнены скрипты для скачивания данных с внешних веб-сервисов, с последующей очисткой, обогащением и фильтрацией. Затем на актуализированных данных будет обучена модель.

Процедура скачивания и обработки данных, с последующим обучением модели, будет также выполняться ежедневно в 02:00.

Примечание. Данная процедура находится в состоянии бета-тестирования.

В результате в контейнеры будут сохранены данные для предсказания индекса качества воздуха, обученная модель и последние метрики.

Затем будет на внутреннем порту 8041 будет запущен веб-сервис.