МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное   
образовательное учреждение высшего образования  
«Самарский национальный исследовательский университет   
имени академика С.П. Королева»

(Самарский университет)

Институт информатики и кибернетики

Факультет информатики  
Кафедра технической кибернетики

**Отчет по лабораторной работе №2**

Дисциплина: «Инженерия данных»

Тема: **«Инференс и обучение НС»**

Выполнил: Мелешенко И.С.

Группа: 6233-010402D

Самара 2023

# Часть 1. Построение пайплайн для инференса данных.

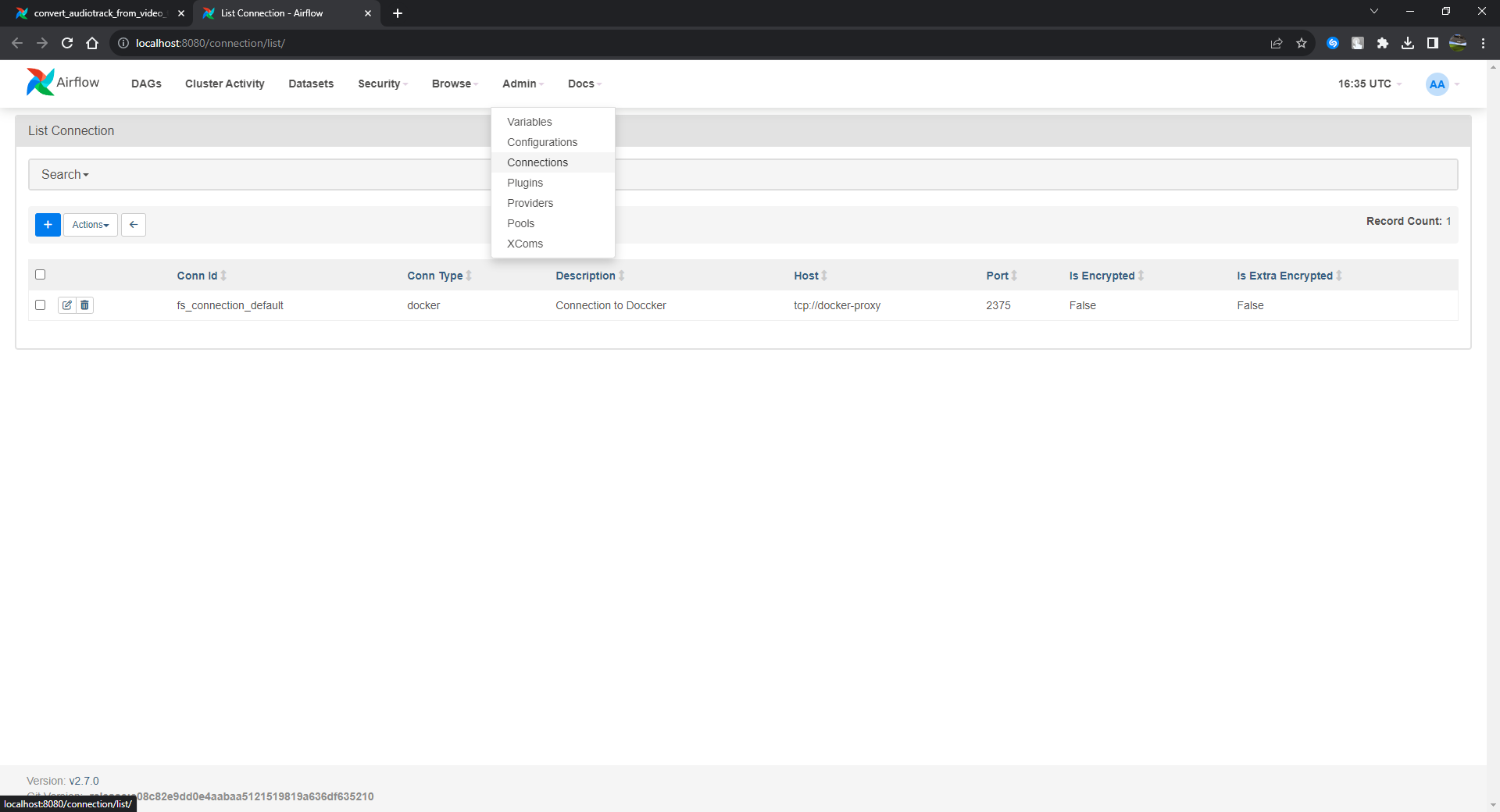
## Шаг 1. Разработка и реализация DAG-a

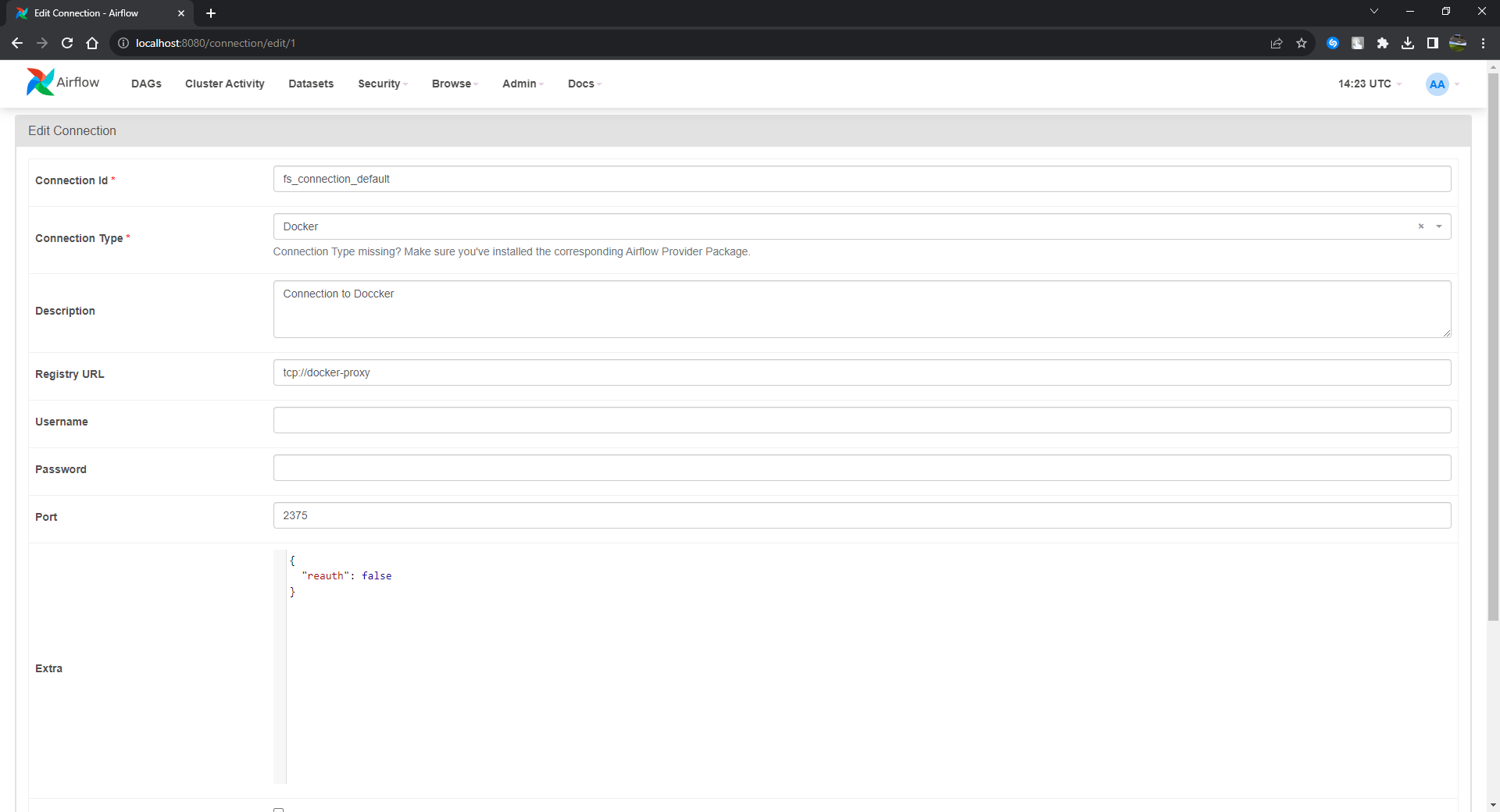
В рамках первого задания необходимо реализовать пайплайн, который реализует систему "Автоматического распознавания речи" для видеофайлов. Построенный пайплайн будет выполнять следующие действия поочередно:

* Производить мониторинг целевой папки на предмет появления новых видеофайлов.
* Извлекать аудиодорожку из исходного видеофайла.
* Преобразовывать аудиодорожку в текст с помощью нейросетевой модели.
* Формировать конспект на основе полученного текста.
* Формировать выходной .pdf файл с конспектом.

Для реализации описанных действий мы будем использовать DockerOperator, а также FileSensor для получения необходимого видеофайла.

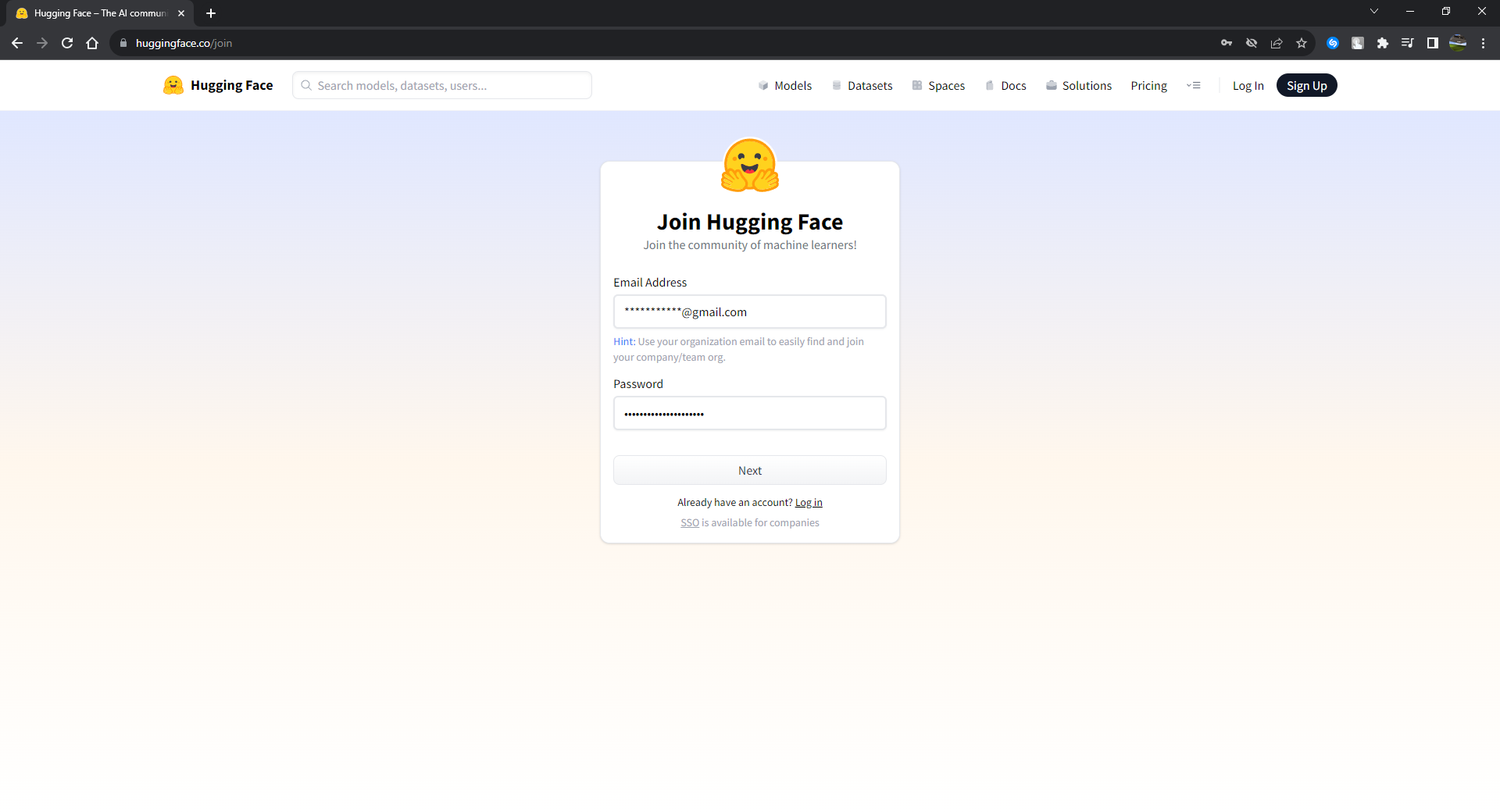
Для работы task-a по ожиданию получения нового видео необходимо создать новое подключение к airflow. Для создания подключения переходим в Airflow по адресу <http://localhost:8080/connection/list/> или мы можем в Airflow пройти по пути Admin>>Connections, как на рисунке ниже.

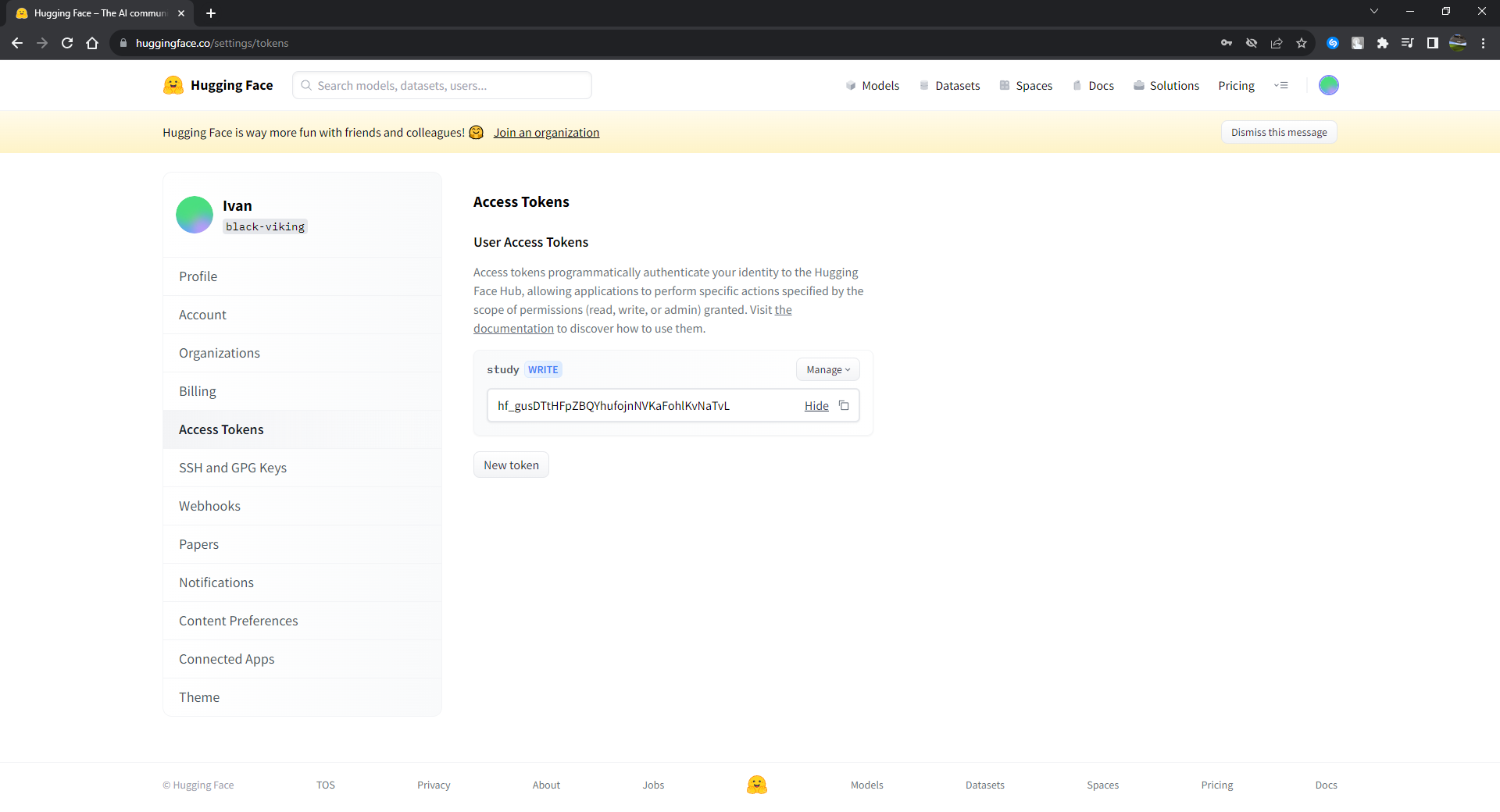
  
Рисунок 1 – Создание Connection

  
Рисунок 2 – Параметры Connection

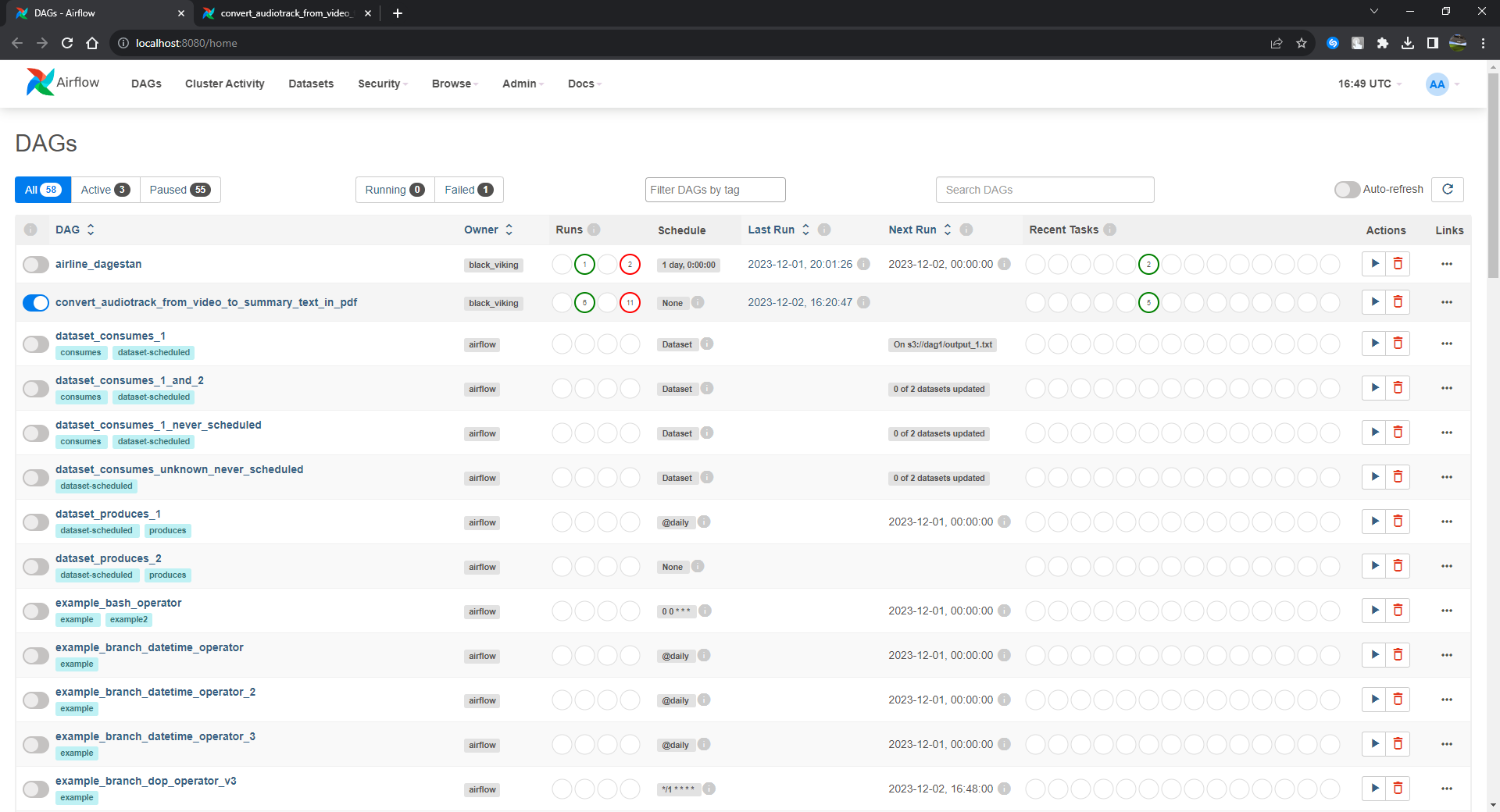
## Шаг 2. Регистрация на huggingface и получения токена API.

Далее для того, чтобы можно было преобразовать наш аудиофайл в текст, а после получить из него summary, необходимо зарегистрироваться на <https://huggingface.co/> и получить токен API с правами записи для возможности посылки и получения запросов к сайту.

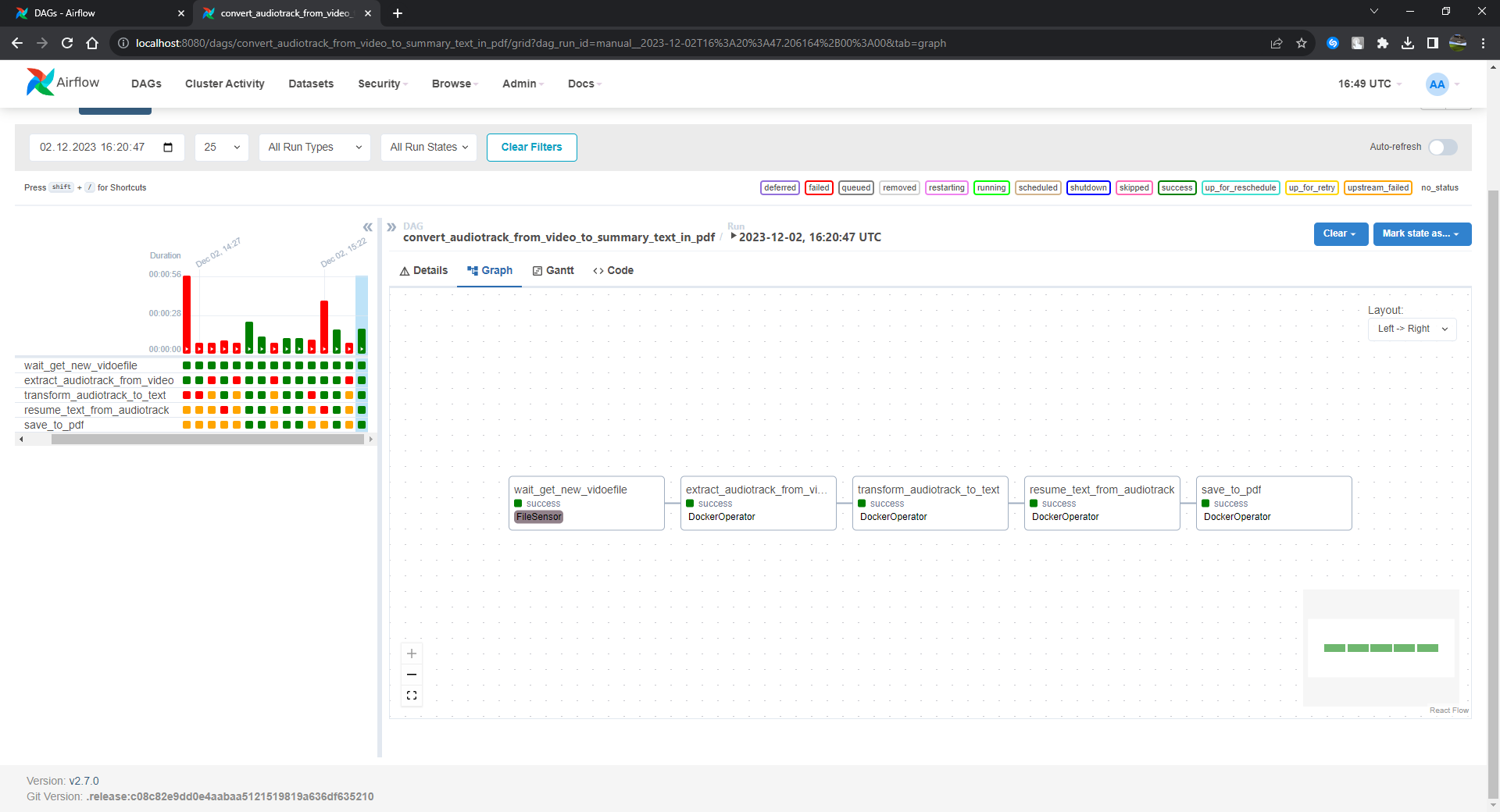
  
Рисунок 3 – Регистрация на huggingface

  
Рисунок 4 – Получение токена API

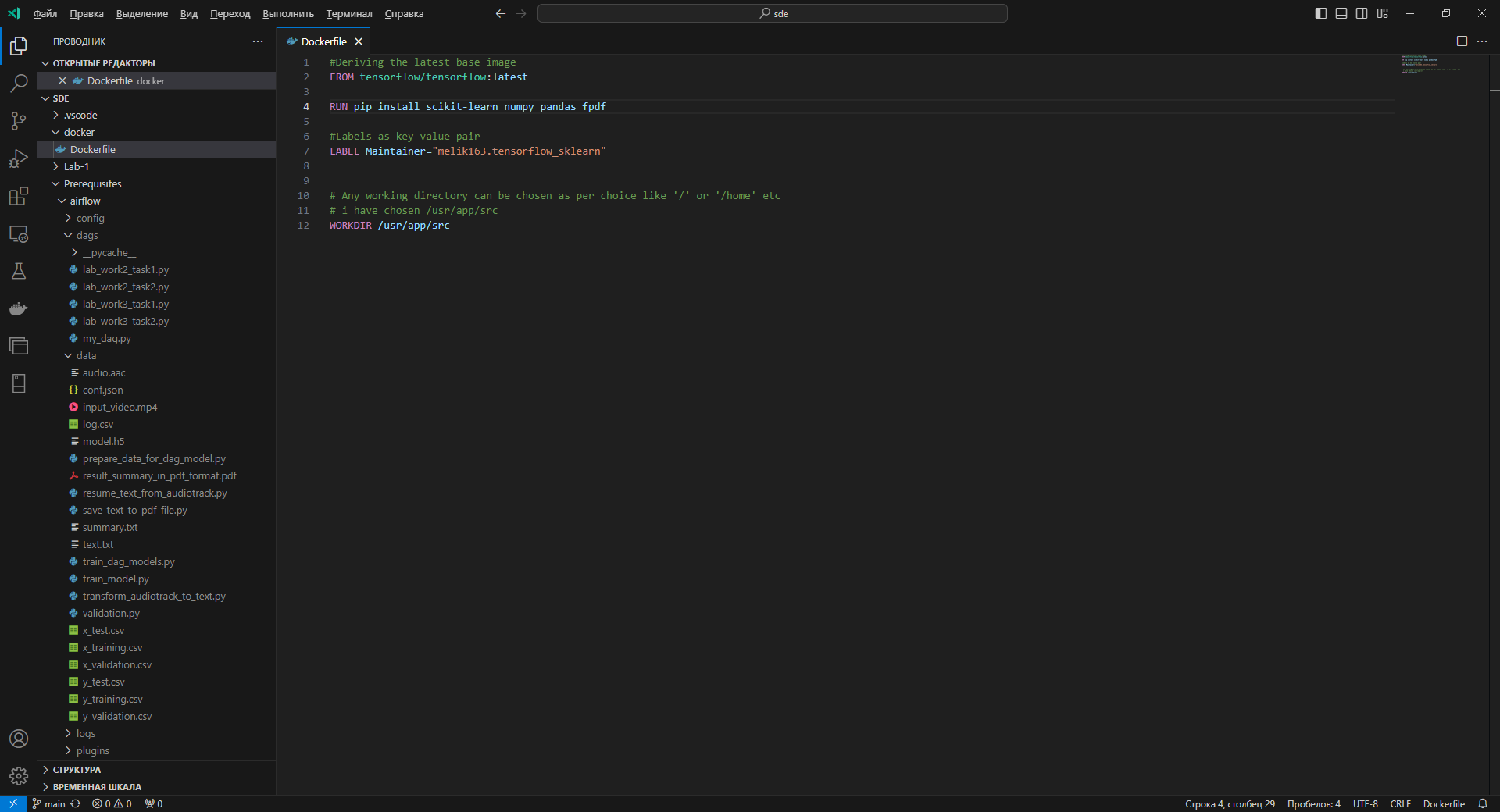
Теперь после всех необходимых настроек можем запустить наш DAG. Переходим в airflow: <http://localhost:8080/home>

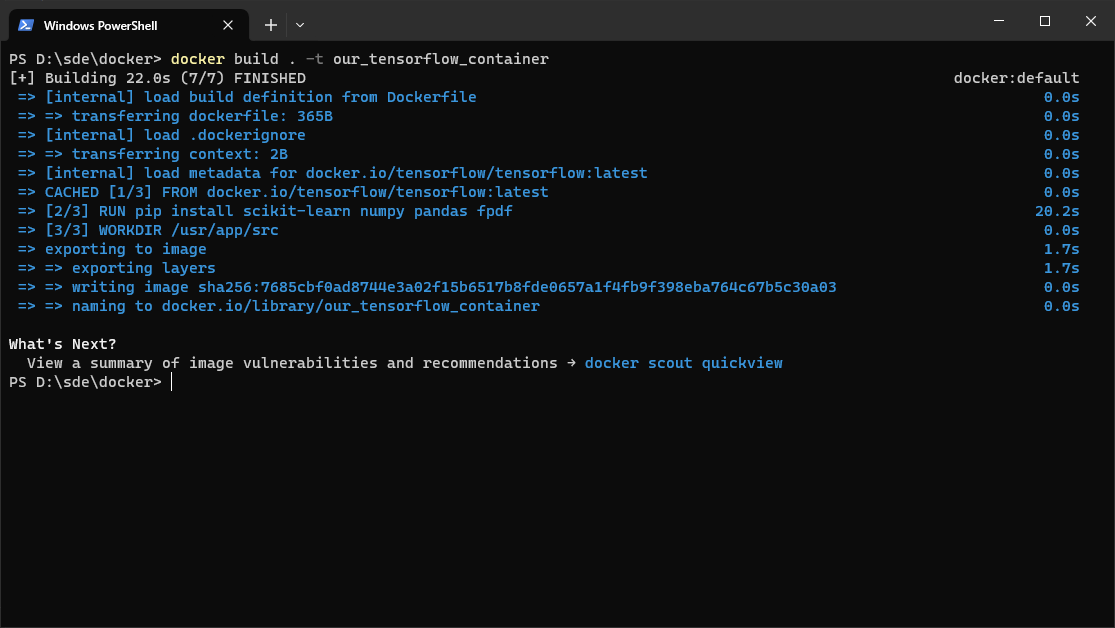
  
Рисунок 5 – Поиск DAG-a.

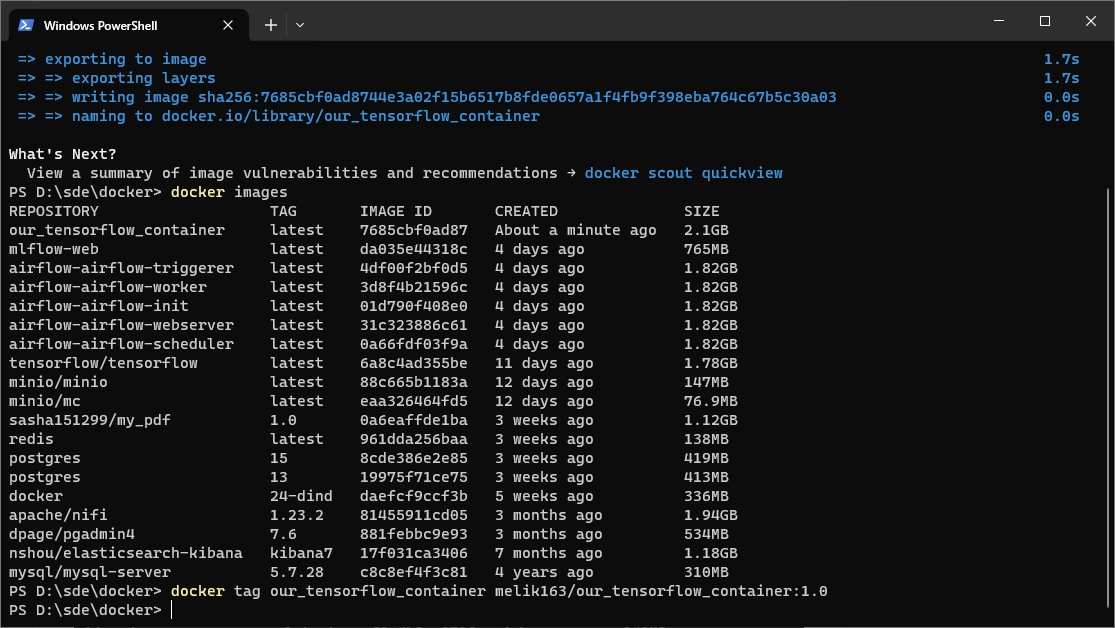
Далее запускаем наш DAG и наслаждаемся процессом.

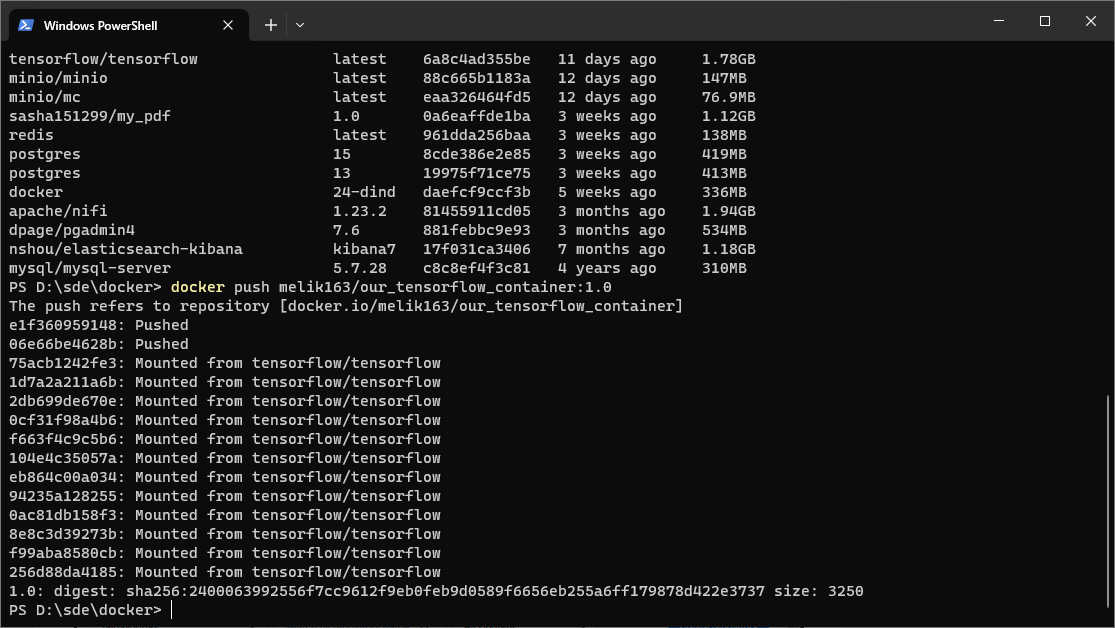
  
Рисунок 6 – Запуск DAG-a.

Для сохранения конспекта в PDF, необходимо было использовать библиотеку fpdf. Создадим необходмый для этого образ в Docker. Процесс представлен ниже.

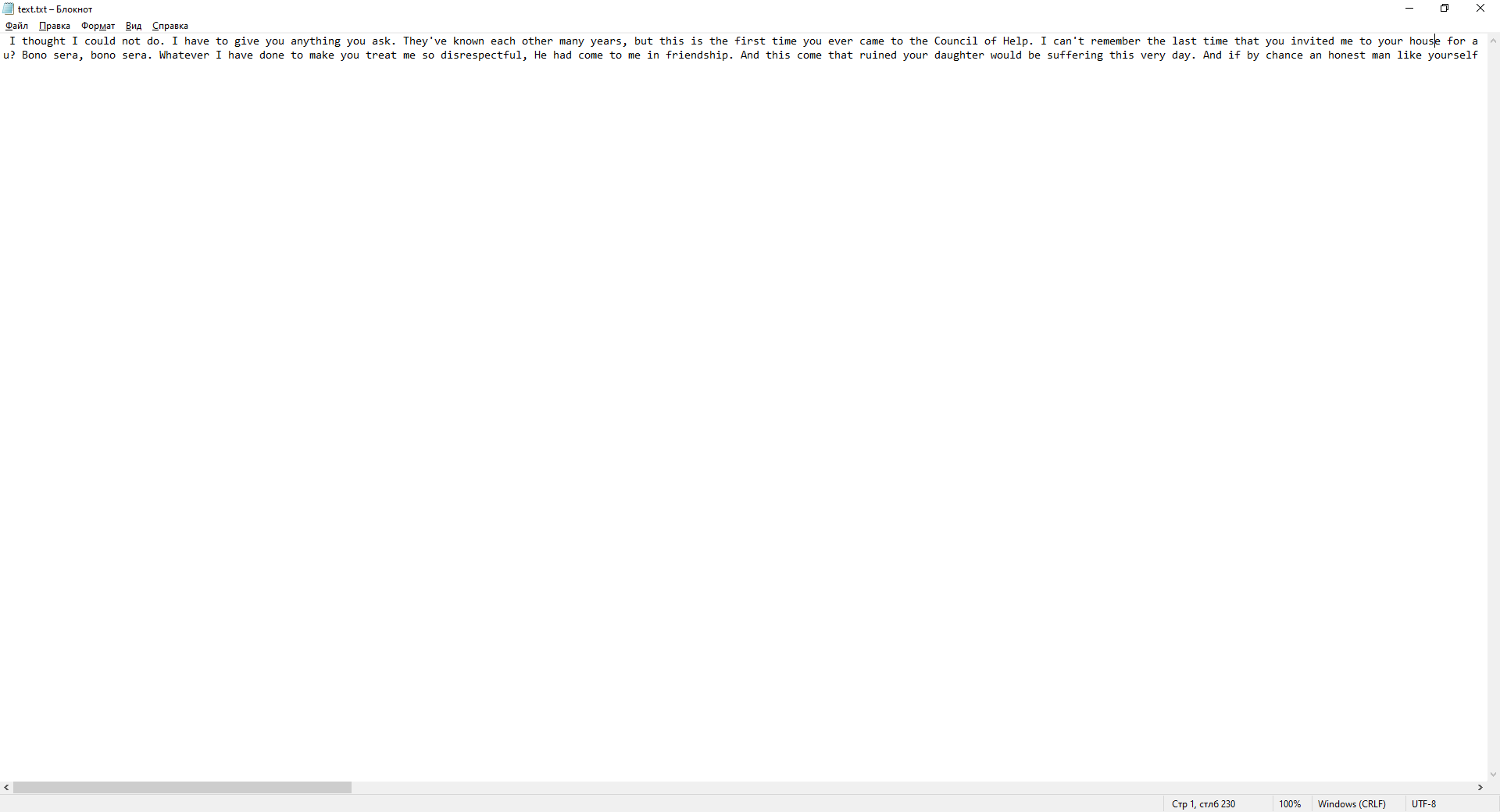
  
Рисунок 7 – Создание Dokerfile

  
Рисунок 8 – Сборка образа

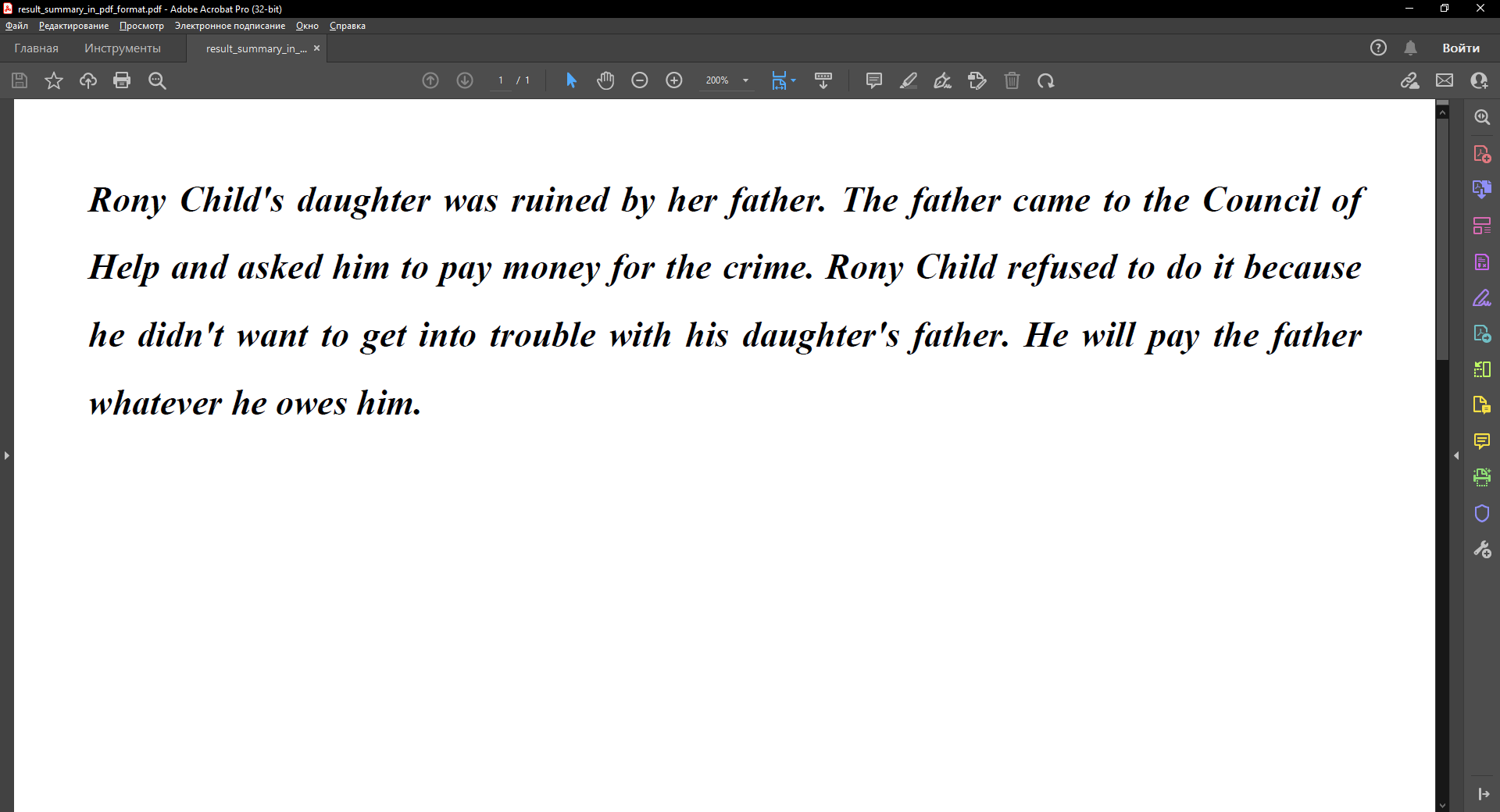
  
Рисунок 9 – Присвоение тега образу

  
Рисунок 10 – Отправка образа в DockerHub

В качестве иходного видео использовался фрагмент из кинофильма «Крестный отец» длительностью 3 минуты 11 секунд. После чего мы получали аудиодорожку, которая использовалась в качетсве основы для получения текстового файла.

  
Рисунок 11 – Результат работы huggingface по преобразованию аудио в текст

После чего полученный результат мы еще раз передавали huggingface для получения уже конспекта по отправленному нами файлу. Полученный результат записывали pdf-файл.

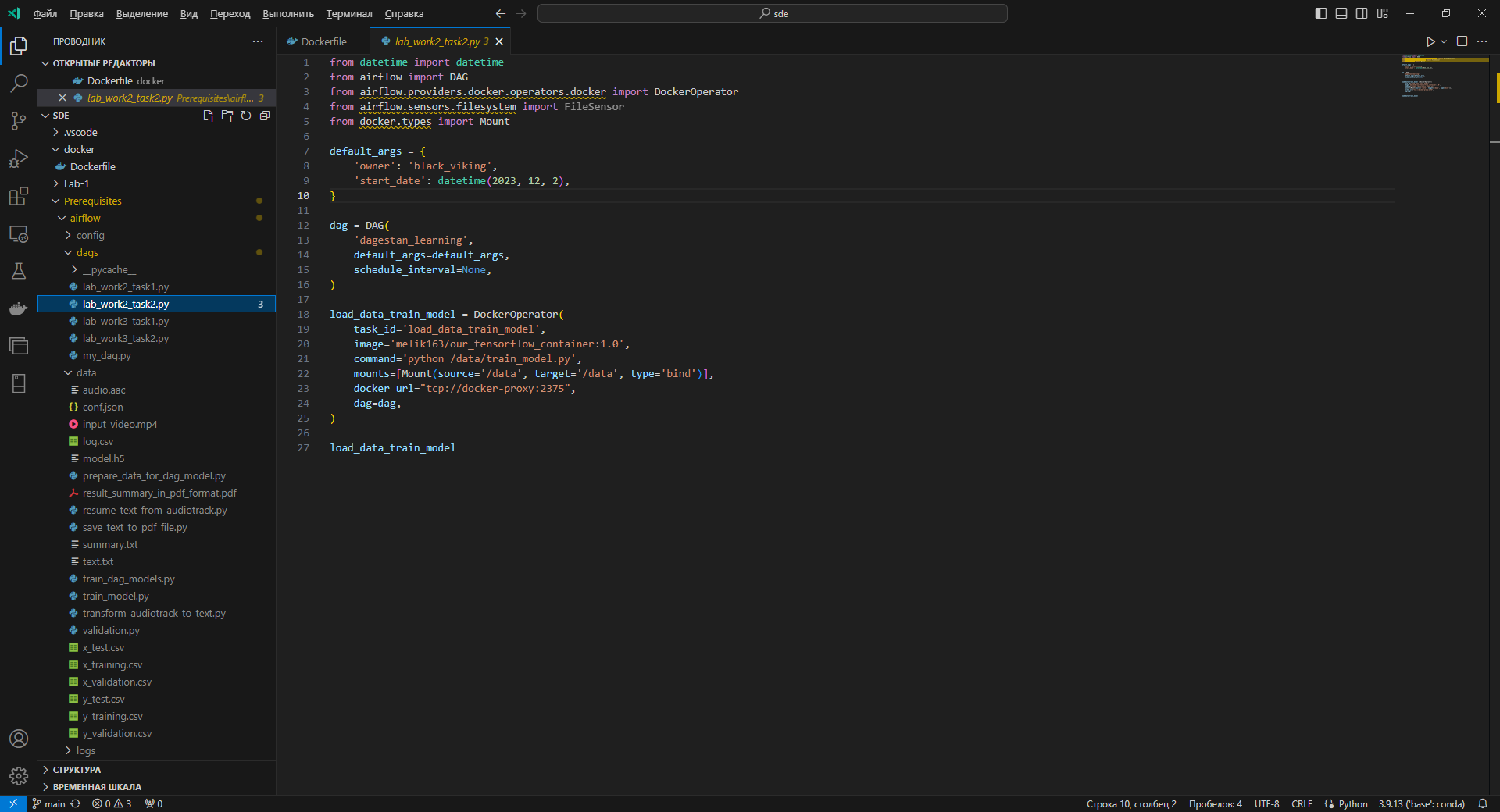
  
Рисунок 12 – Конспект текстового файла.

Получилось неплохо. Перейдем ко второй части.

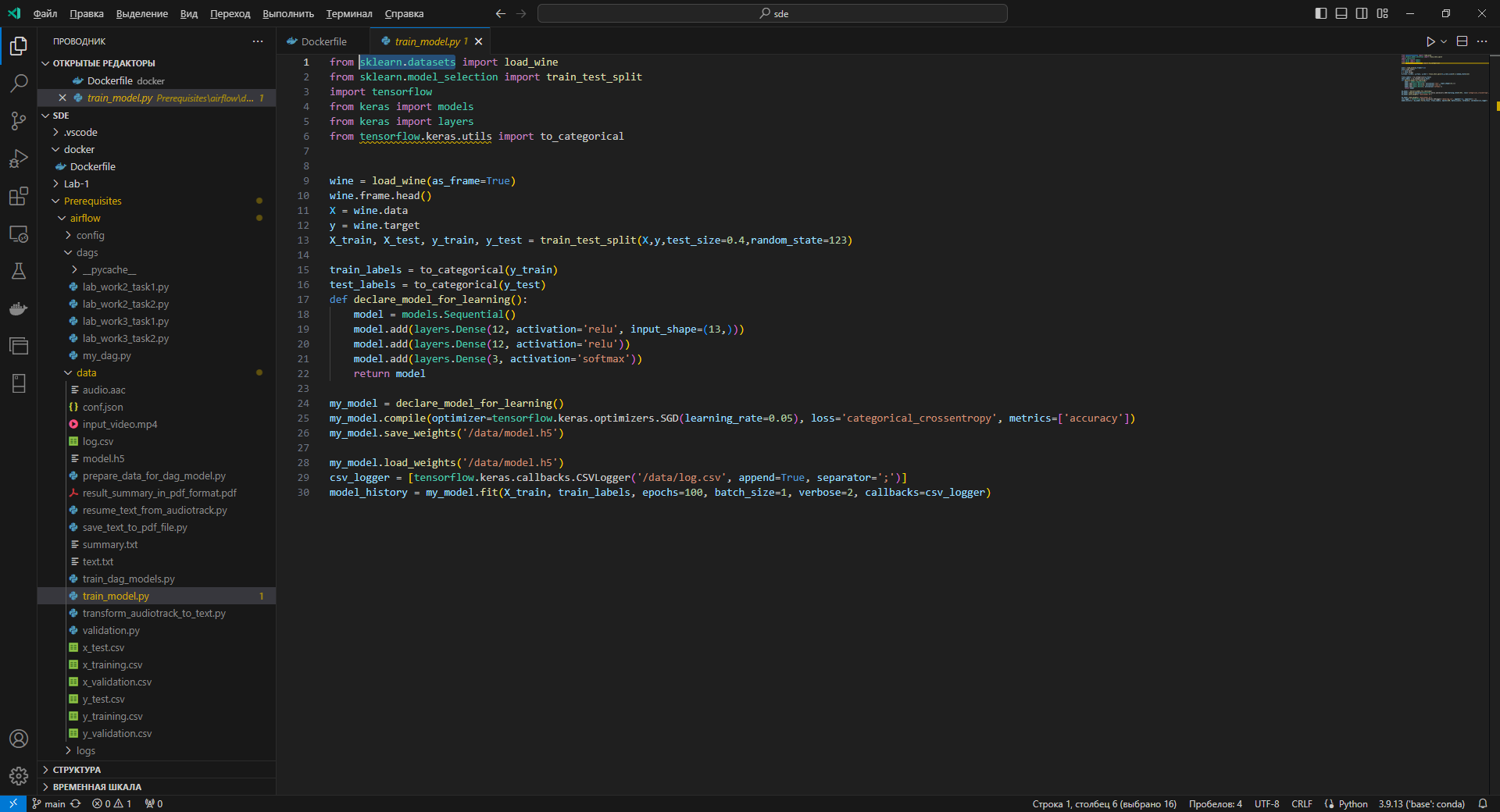
Часть 2. Пайплайн, который реализует систему автоматического обучения/дообучения нейросетевой модели

В рамках второй части лабораторной работы нам необходимо было разработать пайплайн, который реализует систему автоматического обучения/дообучения нейросетевой модели.

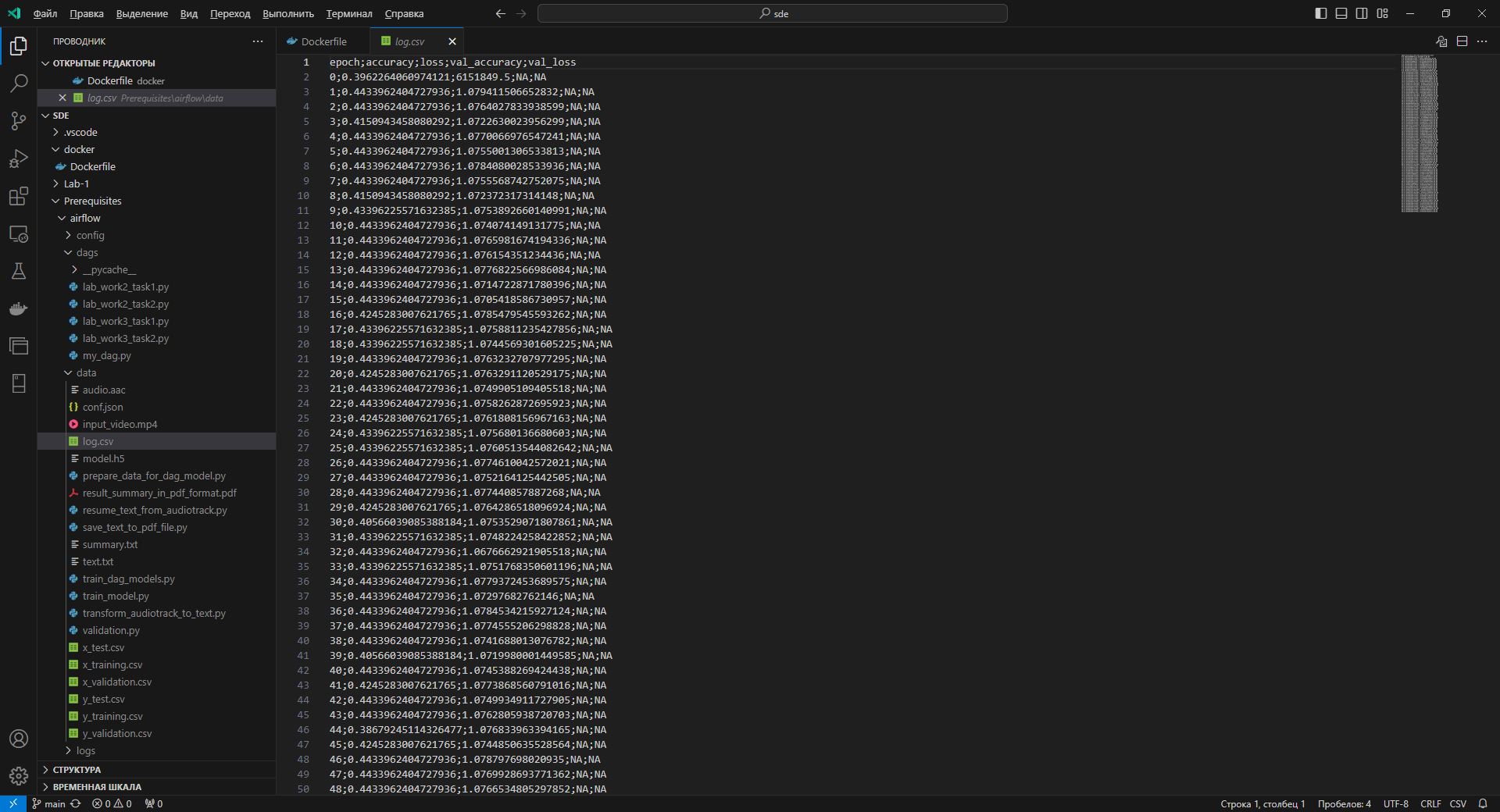
Для этого мы разработали DAG

  
Рисунок 13 - Пайплайн

DAG запускал код, который получал датасет вин load\_wine из sklearn.datasets, после чего мы проводили разбиение данных. Которые передаются в нейросеть, после чего модель проходит обучение. Процесс обучения логируется.

  
Рисунок 14 – Код обучения модели.

В итоге получили вот такой лог обучения.

  
Рисунок 15 – Лог процесса обучения нейросети.

В заключении хотелось бы отметить полезные навыки, полученные в результате выполнения лабораторной работы:

1. Работа с DAG в Airflow
2. Работс сетями на huggingface