Turkmen Whisper Dataset Builder – Instructions

This tool allows you to transcribe `.wav` audio files in the Turkmen language and generate a `metadata.csv` file, ready to be used for Whisper model training. It supports both **offline (local Whisper)** and **online (OpenAI API)** transcription modes.

🔧 System Requirements

- Ubuntu 20.04 or later
- Python 3.8 or newer
- FFmpeg installed
- Internet connection (only required for online mode)

📁 Folder Structure

The files delivered to you include:

```
Whisper_Dataset_Builder/
├── build_dataset.py        → Console application
├── requirements.txt        → Required Python packages
├── .env.example            → Sample for setting OpenAI API key
├── dataset/
│   └── wavs/               → Contains your `.wav` audio files
```

🔑 Step-by-Step Setup Guide (Linux)

1. Open Terminal and navigate to the folder:

```
cd Whisper_Dataset_Builder
```

2. Update your system and install dependencies:

```
sudo apt update
sudo apt install python3 python3-pip ffmpeg
```

3. Create and activate a Python virtual environment:

```
python3 -m venv env
source env/bin/activate
```

4. Install the required Python packages:

```
pip install -r requirements.txt
```

🔐 Set OpenAI API Key (Only if using online mode)

Create a `.env` file in the same folder with the following line:

```
OPENAI_API_KEY=sk-your-api-key-here
```

You can also export it temporarily in terminal:

```
export OPENAI_API_KEY=sk-your-api-key-here
```

🚀 Running the Application

▶ Offline Mode (Default) – uses local Whisper model

```
python build_dataset.py --model base
```

You can change the model to: `tiny`, `base`, `small`, `medium`, `large`

▶ Online Mode – uses OpenAI transcription API

```
python build_dataset.py --online
```

Note: Online mode requires your OpenAI API key set in `.env`

📩 Output

After execution, a file named `metadata.csv` will be created in the same folder. It will contain two columns:

```
file,text

common_voice_tk_36921853.wav, Se có adevilin want kavam b traffak globis

common_voice_tk_36921854.wav," Yaşadım, yaşadım, ölürün."
```

This file can be directly used for fine-tuning Whisper models using Hugging Face.

🆘 Support

If you face any errors during setup or usage, feel free to reach out to me. I will be happy to help you troubleshoot or set things up.