Université de Montréal

IFT-6758-B - A22 - Data Science

Homework 2

Homework Scoring:

Section	Score
Question 1 (.py)	22.5
Question 1 (notebook)	15
Question 2 (.py)	20
Question 3 (.py)	27.5
Question 4 (notebook)	15

In general your homework will be autograded, i.e. you must **not** modify the signature of the defined functions (same inputs and outputs).

Background

The goal of this homework is to get some experience building a dataset from online sources, cleaning it and performing some basic visualizations. To do so, you will make use of scraping libraries, regex and `matplotlib`.

Specifically, we will be working with an audio dataset provided by Google Research described here. However, we will assume it is provided in somewhat of a raw format: a CSV file with YouTube IDs, timestamps and tags.

It will be your job to take this CSV, download the associated audio, format them to only include the relevant segment and clean up the data so it can be used for some downstream ML task (for example, training an audio classifier or a generative model to create similar audio samples).

In addition, you will be visualizing various aspects of the dataset to gain a better understanding of the dataset.

Getting started

This assignment has 2 components that need to be completed: - The `.py` files contain functions that should be filled in as specified in the comments - The `visualization.ipynb` contains cells that need to be filled in and run as well as some basic questions that should be answered in markdown cells.

Start by setting up a virtual environment as you did in the previous homework and install the contents of `requirements.txt`.

It is then recommended to complete the questions in order (starting from the `.py` file then the corresponding sections of the notebook) as the output of previous questions is sometimes used for later questions.

Questions

1. Understanding and visualizing the dataset

To begin, we want to make `audio_segments.csv` more human readable and better understand the distribution of labels. Complete the functions in `q1.py` then fill in and run the cells in `visualization.ipynb` under the section **Question 1**.

Looking at `audio_segments.csv`, the labels for each video correspond to the ID of the label and not the actual label name.

2. Downloading and processing data

Now that we've cleaned up the `.csv`, it's time to get at what we're really interested in: the audio. Complete the functions in `q2.py` which will be the building blocks of our small data processing pipeline. One function should download the audio and the other should cut it to only include the segment mentioned in the `.csv`

3. Building the dataset using a pipeline

With these building blocks, we will build a very small data pipeline to download and process the entire dataset Complete the functions in `q3.py` then run the cell in `visualize.ipynb`

4. Visualizing audio

Now that the segments are downloaded, complete the cells in `visualize.ipynb` to listen to and visualize some of the audio samples.

References

- https://research.google.com/audioset//download.html
- https://regexone.com/