Université de Montréal

IFT-6758A/3700- Science des Données

Devoir 2

Pointage par question:

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

En général vos devoirs seront notés automatiquement, c'est à dire que vous ne devez **pas** modifier la signature des fonctions définies (mêmes entrées et sorties).

Contexte

L'objectif de ce devoir est d'acquérir de l'expérience dans la création d'un ensemble de données à partir de sources en ligne, de le nettoyer et d'effectuer des visualisations de base. Pour ce faire, vous utiliserez des bibliothèques de scraping, du regex et `matplotlib`.

Plus précisément, nous travaillerons avec un ensemble de données audio fourni par Google Research décrit ici. Cependant, nous supposerons qu'il est fourni dans un format un peu brut : un fichier CSV avec des identifiants YouTube, des timestamps et des étiquettes.

Ce sera votre travail de prendre ce CSV, de télécharger l'audio associé, de le formater pour n'inclure que le segment pertinent et de nettoyer les données afin qu'elles puissent être utilisées pour une tâche ML en aval (par exemple, former un classificateur audio ou un modèle génératif pour créer des échantillons audio similaires).

De plus, vous visualiserez divers aspects de l'ensemble de données pour mieux comprendre l'ensemble de données.

Commencer

Ce devoir comporte 2 volets qui doivent être complétés : - Les fichiers `.py` contiennent des fonctions qui doivent être remplies comme spécifié dans les commentaires - Le notebook `visualization.ipynb` contient des cellules qui doivent être remplies et exécutées.

Commencez par configurer un environnement virtuel comme vous l'avez fait dans les devoirs précédents et installez le contenu de `requirements.txt`.

Il est alors recommandé de compléter les questions dans l'ordre (en commençant par le fichier `.py` puis les sections correspondantes du cahier) car la sortie des questions précédentes est parfois utilisée pour les questions ultérieures.

Questions

1. Comprendre et visualizer la base de données

Pour commencer, nous voulons rendre `audio_segments.csv` plus lisible et mieux comprendre la distribution des étiquettes. Complétez les fonctions dans `q1.py` puis remplissez et exécutez les cellules dans `visualization.ipynb` sous la section **Question 1**.

En regardant `audio_segments.csv`, les étiquettes de chaque vidéo correspondent à l'ID de l'étiquette et non au nom réel de l'étiquette.

2. Télécharger et traiter les données

Maintenant que nous avons nettoyé le `.csv`, il est temps de passer à ce qui nous intéresse vraiment: l'audio. Complétez les fonctions dans `q2.py` qui seront les blocs de construction de notre petit pipeline de traitement de données. Une fonction doit télécharger l'audio et l'autre doit couper l'audio pour n'inclure que le segment mentionné dans le `.csv`

3. Construire l'ensemble de données avec un pipeline

Avec ces blocs de construction, nous allons construire un très petit pipeline de données pour télécharger et traiter l'ensemble des données Complétez les fonctions dans `q3.py` puis exécutez les cellules dans `visualize.ipynb`

4. Visualiser l'audio

Maintenant que les segments sont téléchargés, complétez les cellules dans `visualize.ipynb` pour écouter et visualiser certains des échantillons audio.

Références

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