

## ▼ Scrapping Tool Install

```
!git clone https://github.com/aoifemcdonagh/audioset-processing.git
```

```
Cloning into 'audioset-processing'...
remote: Enumerating objects: 228, done.
remote: Counting objects: 100% (71/71), done.
remote: Compressing objects: 100% (57/57), done.
remote: Total 228 (delta 38), reused 38 (delta 13), pack-reused 157
Receiving objects: 100% (228/228), 12.85 MiB | 14.32 MiB/s, done.
Resolving deltas: 100% (116/116), done.
```

```
%cd audioset-processing/
```

```
/content/audioset-processing/audioset-processing/audioset-processing/audioset-processing
```

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

```
Requirement already satisfied: altgraph==0.17 in /usr/local/lib/python3.7/dist-packages (from -r requirements.txt (line 1))
Requirement already satisfied: future==0.18.2 in /usr/local/lib/python3.7/dist-packages (from -r requirements.txt (line 2))
Requirement already satisfied: pefile==2019.4.18 in /usr/local/lib/python3.7/dist-packages (from -r requirements.txt (line 3))
Requirement already satisfied: youtube-dl in /usr/local/lib/python3.7/dist-packages (from -r requirements.txt (line 4)) (2021
```

## ▼ Scrapping Tool Run

```
!python3 process.py download -c "civil defense siren" #takes a while
```

**Streaming output truncated to the last 5000 lines.**

```
libpostproc 54. 7.100 / 54. 7.100
Input #0, mov,mp4,m4a,3gp,3g2,mj2, from 'https://rr6---sn-qxo7rn7s.googlevideo.com/videoplayback?expire=1647692703&ei=P3clYs'
Metadata:
  major_brand      : dash
  minor_version    : 0
  compatible_brands: iso6mp41
  creation_time    : 2014-06-08T21:07:13.000000Z
Duration: 00:02:46.65, start: 0.000000, bitrate: 128 kb/s
Stream #0:0(und): Audio: aac (LC) (mp4a / 0x6134706D), 44100 Hz, stereo, fltp, 7 kb/s (default)
Metadata:
  creation_time    : 2014-06-08T21:07:13.000000Z
  handler_name     : SoundHandler
Stream mapping:
  Stream #0:0 -> #0:0 (aac (native) -> pcm_s16le (native))
Press [q] to stop, [?] for help
Output #0, wav, to './output/civil defense siren/K0a6WEvNZh0_40.wav':
Metadata:
  major_brand      : dash
  minor_version    : 0
  compatible_brands: iso6mp41
  ISFT             : Lavf57.83.100
Stream #0:0(und): Audio: pcm_s16le ([1][0][0][0] / 0x0001), 16000 Hz, stereo, s16, 512 kb/s (default)
Metadata:
  creation_time    : 2014-06-08T21:07:13.000000Z
  handler_name     : SoundHandler
  encoder          : Lavc57.107.100 pcm_s16le
size= 625kB time=00:00:10.00 bitrate= 512.1kbits/s speed=2.86x
video:0kB audio:625kB subtitle:0kB other streams:0kB global headers:0kB muxing overhead: 0.012187%
ffmpeg version 3.4.8-Ubuntu0.2 Copyright (c) 2000-2020 the FFmpeg developers
  built with gcc 7 (Ubuntu 7.5.0-3ubuntu1-18.04)
  configuration: --prefix=/usr --extra-version=0ubuntu0.2 --toolchain=hardened --libdir=/usr/lib/x86_64-linux-gnu --incdir=/
  libavutil 55. 78.100 / 55. 78.100
  libavcodec 57.107.100 / 57.107.100
  libavformat 57. 83.100 / 57. 83.100
  libavdevice 57. 10.100 / 57. 10.100
  libavfilter 6.107.100 / 6.107.100
  libavresample 3. 7. 0 / 3. 7. 0
  libswscale 4. 8.100 / 4. 8.100
  libswresample 2. 9.100 / 2. 9.100
  libpostproc 54. 7.100 / 54. 7.100
Input #0, matroska,webm, from 'https://rr3---sn-qxoedn7z.googlevideo.com/videoplayback?expire=1647692708&ei=RHclYrnAGaK02_gP'
Metadata:
  encoder          : google
Duration: 00:02:55.82, start: -0.007000, bitrate: 136 kb/s
```

```

Stream #0:0(eng): Audio: opus, 48000 Hz, stereo, fltp (default)
Stream mapping:
  Stream #0:0 -> #0:0 (opus (native) -> pcm_s16le (native))
Press [q] to stop, [?] for help
Output #0, wav, to './output/civil defense siren/K0p_TMUQRK0_0.wav':
Metadata:
  ISFT               : Lavf57.83.100
  Stream #0:0(eng): Audio: pcm_s16le ([1][0][0][0] / 0x0001), 16000 Hz, stereo, s16, 512 kb/s (default)
Metadata:
  encoder            : Lavc57.107.100 pcm_s16le
size=    625kB time=00:00:10.00 bitrate= 512.0kbits/s speed=7.06x

```

```

#change the 'sound definition word' to whatever you just ran the above cell for, ie change all the 'crash' below to screaming, or
!zip -r /content/audioset-processing/output/civil_defense_siren.zip /content/audioset-processing/output/civil_defense_siren/
from google.colab import files
files.download("/content/audioset-processing/output/civil_defense_siren.zip")

```

3/10

```
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
```

## ▼ Uploader

```
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/

from msilib.schema import Directory
import os
import glob

#Download AZCOPY into working directory
#https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10

# SAS Token Fall 2021 for storage account: azureml1765189457 (make sure this is for storage account)
sas = "?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupitfx&se=2022-03-07T06:06:41Z&st=2022-03-06T22:06:41Z&spr=https&sig=bFNwSqXaWyDniC"
blob_url = "https://azureml1765189457.blob.core.windows.net" #change this to the current base blob url
container_name = "arkham-container" #version 3 addition for specific workspace blob store
#DO NOT CHANGE THE LINES ABOVE

def return_subdirectories(directory):
    files = os.listdir(directory)
    folders = []
    # print(files)
    # for file in files:
    #     if os.path.isdir(file):
    #         folders.append(file)
    return files

#CHANGE THE NEXT LINE
main_directory = "" #directory of sub-folders

#DON'T CHANGE THE NEXT ONE THO
directories = return_subdirectories(main_directory) #lists of paths to directory
# print(directories)

#DON'T CHANGE ANYTHING DOWN BELOW
#Made all URLs double quotes (using \") I ran on CMD
def upload_sample(container, file_path, label):
    print(file_path)
    file_name = file_path.split('/')[-1]
    file_name = file_name.split('\\')[1]
    file_name = f"{label}\\{file_name}"
    print(f"uploading {file_name}")
    os.system(
        f"azcopy copy {file_path} \"{blob_url}/{container}/{file_name}{sas}\"")

# Crema flat directory needs sorting
# TESS sorted into emotion directories
def uploader(dirPath, container, label):
    files = glob.glob(main_directory+"\""+dirPath+"\"*")
    for file in files:
        print("processing " + file)
        try:
            upload_sample(container, file, label)
        except Exception as ex:
            print('Exception:')
            print(ex)
            continue

def run():
    i = 0
    for direct in directories:
        print(f"Directory: {direct}")
        uploader(direct, container_name, directories[i])
        i += 1

#Rishabh Singh 2022
run()

adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
adding: content/audioset-processing/audioset-processing/audioset-processing/audioset-processing/output/civil_defense_siren/
```

<https://colab.research.google.com/drive/1tyIQ8gVieGjDPxiLUqZSfo8TeH7u8K8G#scrollTo=dbPF8otfwqip&printMode=true> 5/10

6/10

7/10

Address	Contents	Address	Contents	Address	Contents	Address	Contents	Address	Contents
0000	00	0001	00	0002	00	0003	00	0004	00
0005	00	0006	00	0007	00	0008	00	0009	00
000A	00	000B	00	000C	00	000D	00	000E	00
000F	00	0010	00	0011	00	0012	00	0013	00
0014	00	0015	00	0016	00	0017	00	0018	00
0019	00	001A	00	001B	00	001C	00	001D	00
001E	00	001F	00	0020	00	0021	00	0022	00
0023	00	0024	00	0025	00	0026	00	0027	00
0028	00	0029	00	002A	00	002B	00	002C	00
002D	00	002E	00	002F	00	0030	00	0031	00
0032	00	0033	00	0034	00	0035	00	0036	00
0037	00	0038	00	0039	00	003A	00	003B	00
003C	00	003D	00	003E	00	003F	00	0040	00
0041	00	0042	00	0043	00	0044	00	0045	00
0046	00	0047	00	0048	00	0049	00	004A	00
004B	00	004C	00	004D	00	004E	00	004F	00
0050	00	0051	00	0052	00	0053	00	0054	00
0055	00	0056	00	0057	00	0058	00	0059	00
005A	00	005B	00	005C	00	005D	00	005E	00
005F	00	0060	00	0061	00	0062	00	0063	00
0064	00	0065	00	0066	00	0067	00	0068	00
0069	00	006A	00	006B	00	006C	00	006D	00
006E	00	006F	00	0070	00	0071	00	0072	00
0073	00	0074	00	0075	00	0076	00	0077	00
0078	00	0079	00	007A	00	007B	00	007C	00
007D	00	007E	00	007F	00	0080	00	0081	00
0082	00	0083	00	0084	00	0085	00	0086	00
0087	00	0088	00	0089	00	008A	00	008B	00
008C	00	008D	00	008E	00	008F	00	0090	00
0091	00	0092	00	0093	00	0094	00	0095	00
0096	00	0097	00	0098	00	0099	00	009A	00
009B	00	009C	00	009D	00	009E	00	009F	00
00A0	00	00A1	00	00A2	00	00A3	00	00A4	00
00A5	00	00A6	00	00A7	00	00A8	00	00A9	00
00AA	00	00AB	00	00AC	00	00AD	00	00AE	00
00AF	00	00B0	00	00B1	00	00B2	00	00B3	00
00B4	00	00B5	00	00B6	00	00B7	00	00B8	00
00B9	00	00BA	00	00BB	00	00BC	00	00BD	00
00BE	00	00BF	00	00C0	00	00C1	00	00C2	00
00C3	00	00C4	00	00C5	00	00C6	00	00C7	00
00C8	00	00C9	00	00CA	00	00CB	00	00CC	00
00CD	00	00CE	00	00					



research.google.com/drive/1tuIQ8sVicDDviiUcZ8fc0TzU7w9K0Q#search=Tz=dkDE0atfysin%20riestMede+true 0/10

[illegible]