Script.py

| **import** pandas **as** pd **import** os **import** json **from** pydub **import** AudioSegment **from** moviepy.editor **import** VideoFileClip  # Change these as needed VRT\_FILE\_PATH = '/Users/aayush/Desktop/mid\_eval/2016-01-01\_0000\_US\_CNN\_Erin\_Burnett\_Out\_Front.v4.vrt' VIDEO\_FILE\_PATH = '/Users/aayush/Desktop/mid\_eval/2016-01-01\_0000\_US\_CNN\_Erin\_Burnett\_OutFront.mp4' AUDIO\_FOLDER\_PATH = '/Users/aayush/Desktop/mid\_eval/' JSON\_FILE\_PATH = '/Users/aayush/Desktop/mid\_eval/samples.json'  **def** **main**(): # Check if ffmpeg is installed **if** **not** os.system('ffmpeg -version'): print("ffmpeg not found. Please install it and add to PATH.") **return**  # Check if file paths are valid **if** **not** os.path.isfile(VRT\_FILE\_PATH): print(f"VRT file not found: {VRT\_FILE\_PATH}") **return** **if** **not** os.path.isfile(VIDEO\_FILE\_PATH): print(f"Video file not found: {VIDEO\_FILE\_PATH}") **return**  **try**: vrt\_data = pd.read\_csv(VRT\_FILE\_PATH, sep="\t", header=**None**) **except** pd.errors.ParserError **as** e: print(f"Error while reading VRT file: {e}") **return**  vrt\_data.columns = ["text", "start", "end"] vrt\_data.to\_csv('segments.txt', sep='\t', index=**False**)  **try**: clip = VideoFileClip(VIDEO\_FILE\_PATH) clip.audio.write\_audiofile("audio.mp3") **except** Exception **as** e: print(f"Error while processing video file: {e}") **return**  audio = AudioSegment.from\_mp3("audio.mp3")  samples = []  **for** index, row **in** vrt\_data.iterrows(): **try**: start\_time = int(float(row['start']) \* 1000) # pydub works in milliseconds end\_time = int(float(row['end']) \* 1000) segment\_audio = audio[start\_time:end\_time] filename = AUDIO\_FOLDER\_PATH + f"audio\_{index}.mp3" segment\_audio.export(filename, format="mp3")  # Append the information to samples sample\_info = { 'audiofilepath': filename, 'text': row['text'], 'duration': row['end'] - row['start'] } samples.append(sample\_info) **except** Exception **as** e: print(f"Error while processing row {index}: {e}")  **with** open(JSON\_FILE\_PATH, 'w') **as** json\_file: json.dump(samples, json\_file)  **if** \_\_name\_\_ == '\_\_main\_\_': main() |
| --- |

Convert vrt to txt

| **from** bs4 **import** BeautifulSoup  **def** **process\_vrt\_file**(file\_path):  **with** open(file\_path, 'r') **as** file:  data = file.read()   soup = BeautifulSoup(data, 'xml')   **with** open('output.txt', 'w') **as** out\_file: # Open output file for writing  **for** turn **in** soup.find\_all('turn'):  sentences = turn.find\_all('s')  **for** sentence **in** sentences:  words = []  **for** word **in** sentence.find\_all(**True**, recursive=**False**): # direct children only  **if** word.name **not** **in** ["meta", "Unknown"]: # not interested in these tags  words.append(word.text)  sentence\_text = " ".join(words)  out\_file.write(sentence\_text + '\n') # Write sentence to output file with newline  # Use the function process\_vrt\_file('/Users/aayush/Desktop/mid\_eval/2016-01-01\_0000\_US\_CNN\_Erin\_Burnett\_Out\_Front.v4.vrt') |
| --- |