Dataset Name

English Speech-Keywords MultiModal Dataset

Description

This dataset contains a collection of speeches in English, along with the corresponding main keywords and concepts discussed in each speech. The speeches were collected from various sources, including political speeches, business presentations, and academic lectures. Pre-trained keyword extractor models automatically extracted the keywords and concepts from the speeches.

Source

The primary source of this dataset is from **LibriSpeech: Small Dataset** in Kaggle. This dataset contains about 45K audios with their transcripts as well. We use and select 3k of them which is more relevant to our works. In addition, we extract waveforms from audio files and keywords from transcripts using pre-trained models. We could select more than 10K of this dataset and make a larger final dataset but because of resource limitations in RAM and GPU, we just select this 3k now. We planned to make this dataset larger in the future. In the next steps, we also extract BERT embedding of keywords and wave2vec embedding of audio for each speech in our dataset.

Variables

This dataset splits into 4 parts. Training, validation, test, test with negative samples. You can see them in the corresponding Google Drive link. Each of them is a list and you can load them in python easily with the pickle built-in library. Each item on the list is a dictionary that contains these variables inside it:

id: A unique identifier for each item.
audio_waveform: The audio waveform of the speech.
keywords: A list of keywords discussed in the speech.
bert_embedding: A BERT embedding of keywords.
audio_embedding: A facebook/wav2vec2 embedding of audio.

A dataset **test with negative samples** has two more keys inside each item:

candidates: A list of 6 candidates' ids of audios for corresponding keywords. one of them is related to this item and the others are not related.
label: Index of related text inside candidates.

Usage

This dataset can be used for a variety of research questions related to speech analysis, natural language processing, and topic modeling. We gonna use it in a multimodal model which gonna creates a joint embedding space for keywords and concepts of the speech with speech. So for that model, we need keywords and audio of speeches. Thus we create this dataset.