Dataset Name

Persian Speech-Transcript MultiModal Dataset

Description

This dataset contains a collection of speeches in Persian, along with the corresponding main transcripts, keywords, and concepts discussed in each speech. The speeches were collected from Persian podcasts. Speech Segmentation models are used to convert each episode of podcasts to meaningful long enough chunks. After that, Persian ASR models are used to extract transcription from audio. Then Pre-trained Persian keyword extractor models are used to create keywords for each chunk transcription.

Source

The primary source of this dataset is **Radiomarz** podcasts. This dataset contains more than 70 hours of podcast audio with their transcripts as well. As mentioned before, we create an automatic keyword extraction pipeline to have a dataset of audios with their concepts and keywords for different meaningful chunks of data.

Variables

You can access this dataset from this Google Drive link. A dataset is a dictionary that the values are podcast episode names and in each episode, we have a list of items that each item is a chunk of audio. In each chunk you have these variables:

start_time: Start time of that chunk.
end_time: End time of that chunk.
keywords: A list of keywords discussed in this chunk.
transcription: Transcript of that chunk.

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.