Steps to find the number of Unique Speakers and classify the speakers into male and female

1. Youtube Video Link

https://www.youtube.com/watch?v=yX5EJf4R77s

2. Extracting Audio from Video Using Pytube

3. Converting .mp3 to .wav format

The deep learning model which gave the number of unique speakers needs the audio in .wav format

4. Diarization Using Pynote from Hugging Face

https://huggingface.co/pyannote/speaker-diarization

Speaker diarization is the process of partitioning an audio stream into segments according to the identity of the speaker. The goal is to determine "who spoke when" in a given audio recording.

Output on our audio:

```
Number of unique speakers: 4
Speaker: SPEAKER_01, Start: 1.17846875, End: 8.89034375
Speaker: SPEAKER_00, Start: 8.89034375, End: 39.400343750000005
Speaker: SPEAKER_02, Start: 39.400343750000005, End: 51.533468750000004
Speaker: SPEAKER_00, Start: 51.533468750000004, End: 57.69284375
Speaker: SPEAKER_03, Start: 57.659093750000004, End: 62.87346875
Speaker: SPEAKER_02, Start: 62.83971875, End: 64.03784375000001
Speaker: SPEAKER_03, Start: 63.98721875000001, End: 64.84784375000001
Speaker: SPEAKER_01, Start: 64.84784375000001, End: 77.87534375
Speaker: SPEAKER_03, Start: 73.70721875000001, End: 73.72409375000001
Speaker: SPEAKER_03, Start: 77.97659375, End: 82.36409375000001
Speaker: SPEAKER_01, Start: 78.73596875, End: 79.10721875
Speaker: SPEAKER_01, Start: 81.03096875, End: 86.53221875
```

```
Speaker: SPEAKER_00, Start: 83.30909375, End: 90.53159375
Speaker: SPEAKER_01, Start: 90.53159375, End: 93.87284375
Speaker: SPEAKER_00, Start: 93.87284375, End: 100.13346875
Speaker: SPEAKER_02, Start: 97.28159375, End: 101.98971875000001
Speaker: SPEAKER_01, Start: 101.98971875000001, End: 112.13159375000001
Speaker: SPEAKER_03, Start: 112.04721875000001, End: 137.15721875
Speaker: SPEAKER_00, Start: 137.02221875, End: 158.94284375
Speaker: SPEAKER_01, Start: 143.11409375, End: 143.31659375
Speaker: SPEAKER_02, Start: 143.31659375, End: 145.07159375
Speaker: SPEAKER_02, Start: 158.94284375, End: 173.52284375000002
Speaker: SPEAKER_01, Start: 173.52284375000002, End: 182.31471875000003
Speaker: SPEAKER_00, Start: 181.84221875, End: 186.19596875000002
Speaker: SPEAKER_01, Start: 183.88409375, End: 186.09471875
```

5. Used LLM to identify the maximum continuous speaking duration for each speaker

6. Based on the above time stamp, I can crop the original audio for every speaker

7. These audios can be given to another deep learning model from Hugging Face that can do the gender recognition and give the count of male and female speakers

https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech