COP290 Assignment - 2 Subtask - 1

Yash Bansal
2022CS511333

1 Problem Statement

1. Transcribe videos and audios to there transcipt in text format, which can thereby be used for various purposes..
2. Parse pdf files to generate selectable as well as unselectable text from them in a suitable text format.

2 Libraries and Engines used

2.1 Moviepy editor python library

- This is a python library used for various videos and audio processing purposes. In this project, it is used specifically for extracting audio from a video file in .mp3 or .wav format.
- These specific formats are used as these audio files will be then transcibed, which is best done in .mp3 or .wav format. Also, this library is used for cropping the audio and video files, in case complete transciption is not required.
- Link for the library: https://pypi.org/project/moviepy/

2.2 Whisper OpenAI library

- This is a open source library managed by OpenAI. It uses pytorch (another machine learning library) for its base.
- This library can transcribe multiple languages, and produces very good results. But it is somewhat slower compared to other libraries discussed below. It supports .mp3 audio format.
- Link for the library :- https://pypi.org/project/openai-whisper/

2.3 Speech recognition library

- This is a open source library. Its google speech recognizer is used for transcribing audio. It supports .wav audio format.
- This library does not give very good results.
- Link for the library :- https://pypi.org/project/SpeechRecognition/

2.4 AssemblyAI API

- This is a online server which through API call, can transcribe the audio file. It supports .mp3 audio format.
- It gives very good results, as well as is faster compared to other libraries.
- Link for the library :- https://pypi.org/project/assemblyai/

2.5 PyPdf2 Python library

- This is a free, open source library capable of performing multiple pdf operations. It gives very good results, but it able to convert only selectable text in pdf. It cannot convert text in image format inside pdf to text.
- Link for the library :- https://pypi.org/project/PyPDF2/

2.6 PyPdfium2 Python library

- This library is also open source, but gives slower results than pypdf2.
- Link for the library :- https://pypi.org/project/pypdfium2/

2.7 PyMuPdf Python library

- PyMuPDF is a high performance Python library for data extraction, analysis, conversion and manipulation of PDF (and other) documents. It gives very good results.
- Link for the library :- https://pypi.org/project/PyMuPDF/

2.8 Tika Python library

- This library gives very high quality results, which very good formatting of the generated text file, incorporating sections, header/footer and other formattings for the pdf intact, and also generates results very fast. But it is also unable to extract text from images in pdf.
- Link for the library :- https://pypi.org/project/tika/

2.9 Pytesseract Python library

- It is a very good library which uses google's tesseract ocr-engine, and is also able to extract text from images inside the pdf file. But the formatting of generated text is not very good, and it also takes very much time for generate results.
- Link for the library :- https://pypi.org/project/pytesseract/

3 Test cases

The code is tested over multiple test cases of varied lengths of audio, video and pdf files.
Link for the test cases :- https://csciitd-my.sharepoint.com/:f:/g/personal/cs5221133_iitd_ac_in
/Eh2yoFQP5htLnYq3IZb3njMBZbdKJhYuwNI1RSyHPPG4CQ?e=qqoWw7