4 Sprint Plan

Team 27: Abhigyan Ghosh, Tanmay Sinha,Puru Gupta, Sriven Reddy

Sprint 1

1. **Identification:**
    1. Get a list of websites
    2. Check if API available
    3. Check if transcript is manually generated
2. **Automation:**
    1. Make a script to download video and transcript
    2. Make a script to feed downloaded video to ffmpeg/libvcodec

Sprint 2

1. **Quality of Data:**
    1. Include data with multiple speakers, background noise
    2. Data should have Indian accent
2. **Database/File System Creation:**
    1. Plan out a database schema and file structure to hold the downloaded data

Sprint 3:

1. **Automation:**
    1. Pass audio to Aeneas for alignment
    2. Try to generalize a script to download from multiple sources
2. **Database Creation:**
    1. Write script to categorize data in the database for easier extraction
3. **Quality of data:**
    1. Scrape enough data for 6000 hours

Sprint 4 (Integration and Test Sprint)

1. **Verification:**
    1. Make a GUI for verification
    2. Select Random Samples to test the database
    3. Make a back end with Finetuneas for human verification
2. **Quality of Data:**
    1. Remove corrupted data