DSCI 510 Final Project Progress Report

Name: Kai Liu
Github Username: kliu9352
USC ID: 3656471371

Project Scope Update

The project scope remains consistent with the proposal ("Linguistic, Prosodic, and Emotional Analysis of TED Talks"). The goal is still to compare 'Science/Tech' vs. 'Arts' communication discourse across the three proposed layers: linguistic, prosodic, and emotional.

Data Sources

My Python project structure has been established, focusing on robust data loading via `src/load.py`. I have successfully implemented methods to acquire all three primary data sources outlined in my proposal.

1. API Data Source (Requirement Met)

- **Source:** YouTube Timed Captions (Source 2)
- **API/Library Used:** I am using the `youtube-transcript-api` Python library, which serves as my primary API data source for this milestone.
- **Data Obtained:** My `src/load.py` module contains a function `get_youtube_transcript(video_id)` that successfully fetches the full transcript list for a given YouTube video ID.
- **Format:** The API returns a list of dictionaries, with each dictionary containing "text", "start", and "duration" keys, precisely as needed for the prosodic rhythm analysis (WPM, pause frequency).
- **Testing:** This API call is verified in `tests.py` with a test case.

2. Other Data Sources (Implemented)

- **Source 1 (Web Scraping):** A function `get_ted_transcript(url)` has been implemented using `requests` and `BeautifulSoup4` to scrape the full-text transcripts directly from TED.com (Source 1).
- **Source 3 (File Load):** A function `load_nrc_lexicon(filepath)` has been implemented using `pandas` to load and parse the NRC Emotion Lexicon (Source 3) into a usable DataFrame for the emotional profiling.

Issues/Difficulties

While data acquisition is successful, I have identified the following potential issues for the next project phases:

- **Web Scraping Fragility (Source 1):** The `get_ted_transcript` function relies on the specific HTML structure of TED.com. If the website's class names or layout changes, the scraper will break. This is a common issue with web scraping.
- **Caption Quality (Source 2):** The `youtube-transcript-api` may pull auto-generated captions if official ones are unavailable. Auto-generated captions often lack proper punctuation and timing, which could skew the "Prosodic Rhythm" (WPM) analysis and require significant pre-processing.
- **Emotional Analysis Context (Source 3):** The NRC Lexicon provides word-level emotion associations (e.g., "rock" = "joy"). This method cannot capture context, negation (e.g., "not happy"), or sarcasm, which may limit the accuracy of the emotional profile.