Statement of Work

Immersive Language Learning With LLMs and AI-Driven Analytics

By Save the Penguins 🐧

Team Members

Christian Aagnes ································ (christianaagnes@g.harvard.edu)
James Cao ······································· (jamescao@g.harvard.edu)
Alyssa Chang ··································· (yujie_chang@g.harvard.edu)

I. Problem Statement

To develop an English language learning application that leverages online media sources to create an immersive learning experience, matching users with content at their proficiency level and allowing users to learn dynamically while engaging with everyday content.

II. Objectives

1. Collect a quality, diverse set of English media sources such as YouTube videos, news articles, podcasts, etc.
2. Create proper transcription of collected English media.
3. Develop a classification model to rate the difficulty of media content and match users with appropriately leveled materials.
4. Provide summaries and translations for users as they engage with content.
5. Implement a scalable backend to handle the deployment of multilingual software to users at different learning levels.
6. Develop an intuitive user interface with integration with media sources, to allow for a smoother, non-intrusive learning experience.

III. Learning Emphasis

This project will use NLP techniques to handle the text/media as well as deep neural networks for our classification model.

IV. Application Mock Design

1) A main dashboard where users can take diagnostics tests, receive content recommendations (based on their language proficiency), track learning progress and review previously encountered words.
2) (Limited) Integration with media source, providing summaries and highlighting key phrases for the user to review while consuming the content.

V. Research and Development

We will read papers and research from linguistics to build our machine learning model (with the correct features) to properly categorize the text into different difficulty levels.

VI. Data Sources

- Training/Test data for ML model: We will be scraping this website (https://learnenglish.britishcouncil.org/) for articles and transcriptions of videos that are already categorized into 5 different difficulty levels. This will serve as the training data for our supervised classification task. Note: we will do a train/test split to evaluate the performance of our model.
- Inference data: In the app we will be showing our users YouTube videos and news articles.

VII. Limitations and Risks

- Lack of accessible and diverse quality sources of media.
- Accuracy of transcription and summarization of material.
- Computational limitations with the integration of complex models.
- Integration and interaction of different media into our application.

VIII. Minimum Components for a Good Project

- Large Data: Diverse and comprehensive sources of media content of varying English language difficulty.
- Scalability: Ability for multiple users to access various parts of the application without compromising performance.
- Complex Models: Use of deep learning models to create a comprehensive rating system for English media and provide users with appropriately leveled content.
- Integration with APIs: Utilize readily available resources such as OpenAI and Google Translate APIs to dynamically interact with new media sources.

IX. Milestones

1) Data Collection and Pre-processing: [10/04/24]
2) Language Model Development and Other NLP Tasks: [10/15/24]
3) Backend Implementation: [10/31/24]
4) Frontend Development: [11/07/24]
5) Integration and API Development: [11/14/24]
6) Final testing and deployment: [12/12/24]

X. Summary

Our language learning application aims to teach English by pulling media from various online sources (e.g., YouTube, podcasts, articles). The content would be transcribed, then categorized by difficulty using our ML model that would take in features like average length of words/sentences, number of punctuation/conjunctions, context of text, etc. Users would first take a diagnostic test to assess their proficiency, and the app would recommend media suited to their level. It would also provide summaries, key vocabulary, and grammar insights for each media piece, generate comprehension questions, and track progress, allowing users to be reranked over time.