ACM AI Team 1: MBTI Personality Project

Team Meeting Time: Every Monday 5:00 - 6:00 PM
Weekly Mentor Meeting: Every Friday 5:00-6:00 PM

Attendees: Kevin, Vanessa, Yashil, Samuel, Chi, Vincent

Important Links:

- **MBTI personality type using tweets :** MBTI Personality Type Twitter Dataset | Kaggle
- **Our github repo: acmucsd-projects/sp23-ai-team-1 (github.com)**
- **Notebook:** https://www.kaggle.com/samuellee1/acm-sp23-team-1-project-mbti
- **Preprocessing**: https://colab.research.google.com/drive/17kLdesMgX6Z6PdiSGeoGPvoFMk-nmP00?usp=sharing

Action Items

- Abandon tutorial approach
- Everyone work on custom dataset for pytorch
- Clean the data and upload cleaned dataset to kaggle (public) (Yashil and Vanessa)
    - Split by '|||' → plot word counts of tweets
    - EDA plot distribution of word counts in tweet
    - Then apply clean (drop foreign languages, remove emojis + regex)
    - Make it more efficient
- Build preprocessing dataset
    - Tokenize text (Samuel)
        - Build dictionary ~20000?
        - Assign words to numbers
    - Pad sequences (Kevin)
        - Find cutoff point for word count (look at distribution)
        - Build and run pad sequence (pad with 0 if under cutoff, cutoff if over)
    - Lemmatization - not needed anymore for deep learning, more for traditional NLP (try if have time Chi)
    - Write up the get item function

- Build model

Summary of Meeting

- Updates of what everyone has done
    - Yashil: removed the emojis
    - Kevin: Gibberish detection, generated box plot for gibberish probability, summarize part of past EDA
    - Vanessa: summarize and generate a pie chart of English and non-English percentages -> we can safely drop non-English test
- Built our own custom dataset

Timeline

- Step 1: Data Cleaning (28th April, Friday of Week 4)
- Step 2: Preprocessing data (1st May, Monday of Week 5)
- Step 3: Model training and validation (8th May, Monday of Week 6)
- Step 4: Decide what kind application to do and start (15th May, Monday of Week 7)
- Step 5: Continuation of the step 4- (Weeks 8,9)
- Have an app ready with Streamlit (Week 9)
- Step 6: Polish it up + presentation + showcase (Week 10)