Title: Immediate MBTI Personality Diagnosis Through Online Posts

Members: Yihang Hu, Jiacheng Yang, Aijia Zhang

Problem Description:

We wish to explore the concept of categorization of personality types using the MBTI test, which is normally an introspective self-report questionnaire that shows people's different psychological preferences when perceiving the world and making decisions. It arises from four dimensions that consist of two categories each. Those dimensions are Extraversion (E) versus Introversion (I), Sensing (S) versus Intuition (N), Thinking (T) versus Feeling (F), and Judging (J) versus Perceiving (P).

In our research, we will focus on constructing an immediate MBTI personality diagnosis through individuals' responses to certain topics. This is accomplished using tokenization for data preparation and multi-variable SVM & natural language processing(NLP) for 16 sub-types categorization. A possible comparison between the two techniques' accuracy might be carried out later. The goal of this project is to not only convey an understanding of the concept but also to seek various implementations and potential improvements to deliver a better result.

Machine Learning Techniques:

We will implement tokenization to pre-process tons of text derived from the personality dataset into standard data and save them into tensor format. Then, we will use a natural language processing(NLP) algorithm like Long short-term memory(LSTM) and multi-variable linear SVM to separate our data into multiple categories. We hope that from this categorization process, we could find some valuable information for identifying people's personalities from simple dialogue, such as their daily conversation. Meanwhile, we want to compare the two different methods(LSTM and linear SVM) of categorization to find out which one produces better accuracy. At last, we will do the training and implement the model to predict a person's personality type based on the person's argument.

Data Description:

link: https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset

This dataset covers 106, 000 records of preprocessed posts and their authors' personality types. The posts are equal-sized: 500 words per sample. The data are first collected from forums like Reddit and PersonalityCafe Forum, compiled together, and then uploaded on Kaggle.

Potential Challenges:

- Implementing all the complex concepts in Python
- Learning new algorithms that are not covered in class, such as NLP and multi-variable SVM
- Finding valid correlations and avoiding potential research error in methodologies