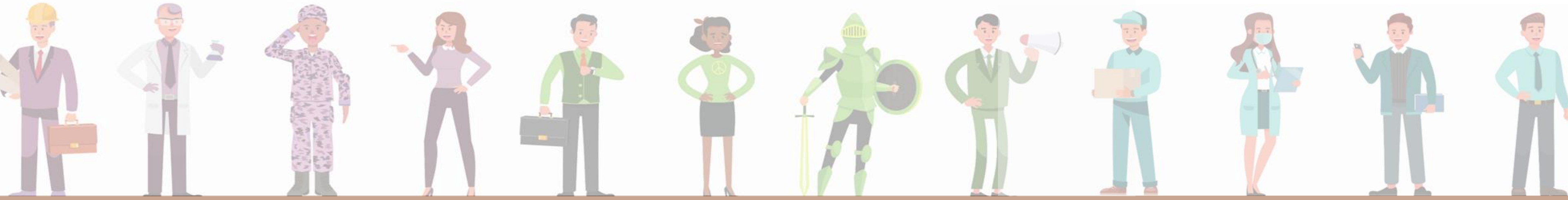
Predicting MBTI Personality Types With Text Data

“You are what you say”

By Jojo Zhou



Motivations

- Knowing your personalities → higher EQ, better decisions, …

- Myers-Briggs Type Indicator (MBTI):
  - Energy -- Extraversion (E) vs. Introversion (I)
  - Information -- Sensing (S) vs. Intuition (N)
  - Decisions -- Thinking (T) vs. Feeling (F)
  - Lifestyle -- Judging (J) vs. Perceiving (P)

- Reasons for a ML model based on texts:
  - Traditional introspective questionnaires are subject to Self-Report Bias
  - Faster, scalable, and practically easier to use

DEMO

http://msia423-1166961.us-east-2.elb.amazonaws.com/

Data Source and RDS Usage

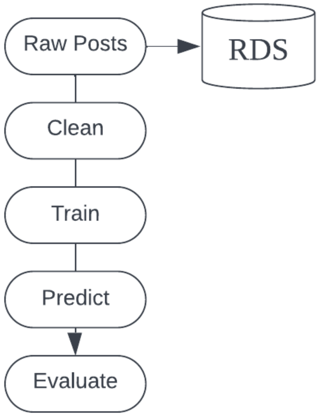
- Kaggle Dataset:
  - (MBTI) Myers-Briggs Personality Type Dataset by MITCHELL J.
  - PersonalityCafe forum, a social media platform that people post and share their MBTI types

- Data Structure
  - 8600 rows of data, one row for one person
  - 2 columns:
    - Type (e.g. INTJ)
    - Posts (last 50 posts separated by "|||", uncleaned)

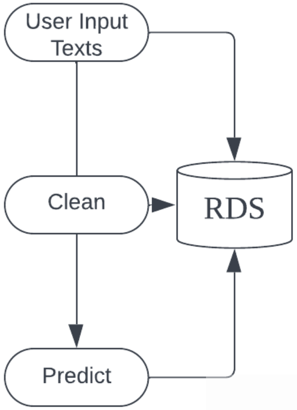
- Data Cleaning
  - Replace url with "link"
  - Remove punctuations
  - Convert to lower cases
  - Lemmatize (NLTK package)

    Stopwords (NLTK package)

- RDS stores:
  - Raw posts data for training
  - Raw user input texts, cleaned texts, and predicted MBTI type



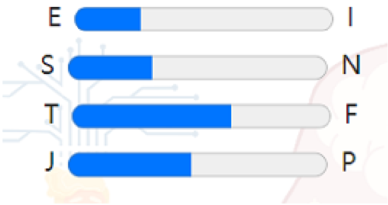
run.py



app.py

Models and Success Metrics

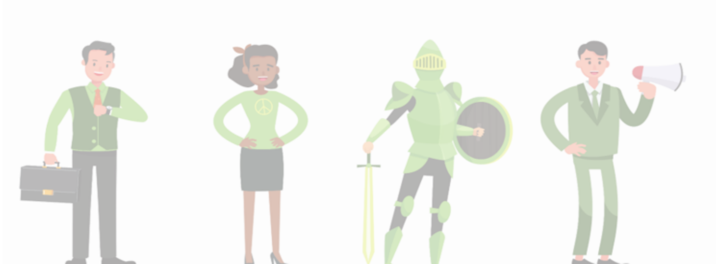
- Models/Approaches
  - TF-IDF vectorizer
    - Max features = 5000
    - Same stopwords as in clean.py
  - Binary Logistic Regression models
    - One model for each dichotomy
      → 4 binary classification models
    - Return 4 pairs of predicted probabilities



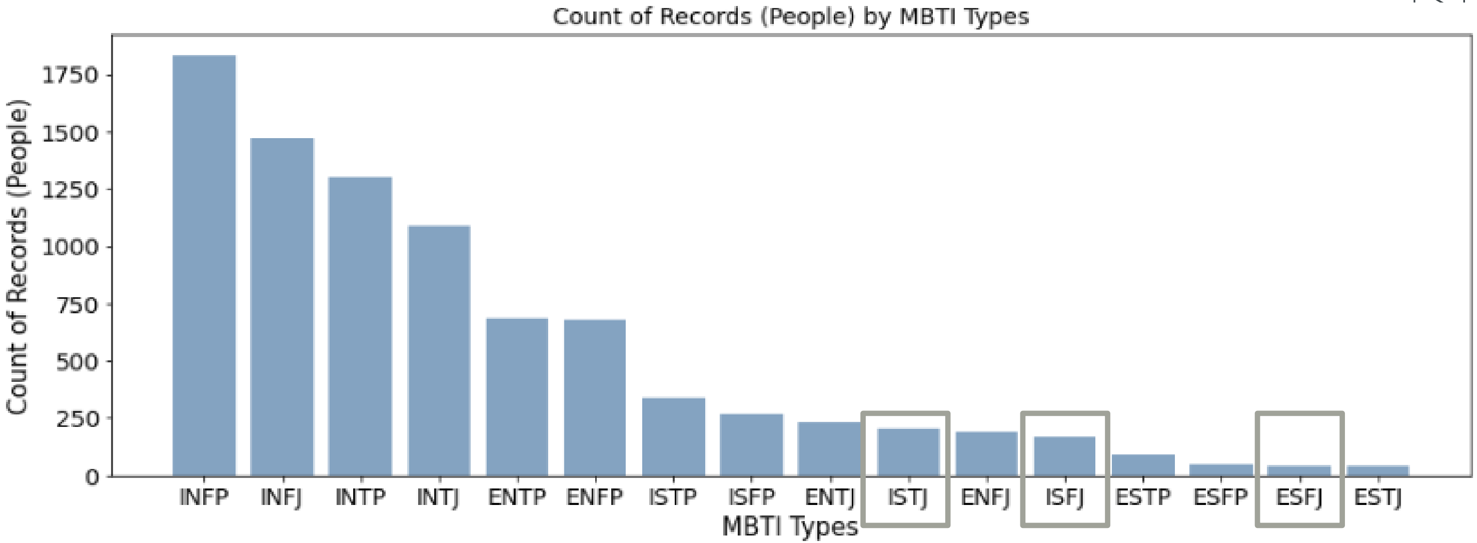
- Online Model
  - Input texts from the web app go through the same cleaning, vectorization, and predicting process as in run.py

- Success Metrics
  - Imbalanced data
    -> accuracy scores are misleading
  - averaged F1-score across all the dichotomous pairs

- Evaluation Results (80-20 train-test split)
  - Previous goal: **0.7**
  - Test result: **0.83**
    - E vs. I – 0.82
    - S vs. N – 0.87
    - T vs. F – 0.86
    - J vs. P – 0.77



Interesting Findings



- Much more I's than E's
- Much much more N's than S's
- Inconsistent with the distribution published by myersbriggs.org between 1972 and 2002, which said ISFJ, ESFJ, and ISTJ are the three most frequent types

Thank you

Do you have any questions?

qingyangzhou2022@u.northwestern.edu

CREDITS: This presentation template was created by **Slidesgo,**
including icons by **Flaticon,** infographics & images by **Freepik**