Frequency Analysis of AI vs. Human Text and Classification Model

COLIN BROGAN

DR. HONG XIAO

Problem & Solution

Problem	Solution
Increase of AI generated Content Becoming more difficult to distinguish	Use word frequency to determine patterns to assist in determining origin. Build classification model to identify between the two.

Project Overview

- MVP: Learning Model to predict origin of written essay text
- What's the goal?
- What does it do?

Why This?

- Interest in AI
- Data Science vs. Game Development
- Resume Building

Languages

Python Anaconda Distribution

- Well Documented
- Common for NLP and Data Science
- Previous Experience

Other Considerations:

- Java
- R
- C++

Environment

Jupyter Notebook

- Cell Structure
- Easy Markdown
- Familiarity

Other Considerations

- Visual Studio
- Pycharm
- Google Prolab

Data Collection

Human Essay Text

- 700 ~ 3000 words
- College Level Submissions
- From Myself, Friends, and Family
- ~ 50 Human Essays

AI Essay Text

- 600 ~ 1300 words
- ChatGPT 3.5 & 4.0
- Given Popular College Essay Prompts
- ~ 70 AI Essays

Feature Selection

Based On Three Primary Areas of Focus

- Stopword Frequency
- POS Tags
  - Nouns
  - Adjectives
  - Pronouns
- Punctuation Usage

adjectives_per_word	lexical_diversity	average_word_length	stopword_frequency	personal_pronouns	possessive_pronouns	label
13.432836	0.265504	5.057032	0.455703	0.020764	0.008029	Human
8.474576	0.394454	4.503578	0.483005	0.044723	0.024150	Human
17.258883	0.331044	5.125343	0.434409	0.024038	0.016484	Human
15.916399	0.348091	5.057271	0.435418	0.024777	0.010561	Human
12.159710	0.336806	4.859375	0.493056	0.039931	0.018229	Human

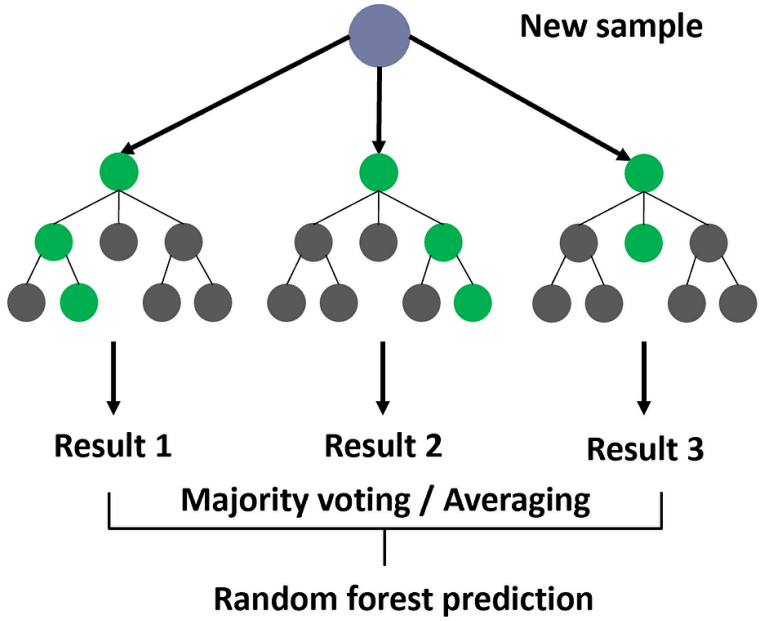
Model Selection

Sci-kit learn Random Forest

- Easier implementation with sci-kit
- Feature Importance Visualization
- Reliable on Relatively Small Datasets

Other considerations

- Support Vector Machine (SVM)



Questions?