You write good: Developing an Algorithm for Automated Scoring of Essays
TF: Yoon Kim

A. "The human mind will always be superior to machines because machines are only tools of human minds."

B. "In order for any work of art—for example, a film, a novel, a poem, or a song—to have merit, it must be understandable to most people."

C. "The well-being of a society is enhanced when many of its people question authority."

Choose one of A, B, C, and write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.

Sound familiar? These are some typical essay prompts that you might receive on standardized tests like the SATs/GREs. Did you know that algorithmic scoring of your essay is a significant part of your final writing score?

In this project, we will be developing machine learning models to automatically score the writing quality of essays---n important social problem with significant practical implications. Through this project you will also become familiar with tools from Natural Language Processing (NLP), from basic techniques such as bag-of-words/bag-of-ngrams and language models, to more advanced techniques such as topic modeling and deep learning (if you so choose to explore them).

Milestones

1. Project Selection: Form a team of 2-3 people

2. Literature Study: This is a rich area of active research with very practical implications. Read two recent papers (2014 onwards) from the following website, and write a 0.5-1 page summary of each.

https://www.ets.org/research/topics/as_nlp/writing_quality/

Focus more on the novelty of techniques rather than other aspects (e.g. the minutiae of data collection/preprocessing).

3. Data Exploration and Cleaning: Much of the data collection has been already done for you. We will be working with a publicy available dataset from the Hewlett Foundation, which can be found here:

https://www.kaggle.com/c/asap-aes

Your first step is to understand the data dictionary and how the data was formatted (e.g. the text was anonymized).

Explore the following, with visualizations where appropriate:
- Essay length vs score
- Vocabulary size (i.e. number of unique words) vs score—how would you take into account misspelled words? Large vocabulary size may mean a sophisticated writer or a novice (who makes lots of spelling mistakes).
- Each essay was scored more than once by an expert. How correlated are the expert judgments? (These are also called as inter-annotator *agreement* and is often a good proxy for how difficult the task is for a human being). Do the correlations vary across essay prompts?

4. Models: Your baseline model will be a simple linear model built on top of bag-of-ngrams features. While ridiculously simple, this is a strong baseline that will be surprisingly hard to beat. You will build two baselines, which are already available in sklearn.

- Naive Bayes model on bag-of-ngrams
- Linear model (either Logistic regression or SVM) on TF-IDF (term-document inverse-document-frequency) bag-of-ngrams features.

From hereon, you will have a choice to explore more sophisticated models, such as:
- topic modeling
- deep learning
- integrating more NLP features (e.g. language model, part-of-speech tags, etc.)