R Senthil Kumar

Automated Essay Scoring

September 20, 2021

Exploratory Data Analysis

The following details were observed while performing EDA:

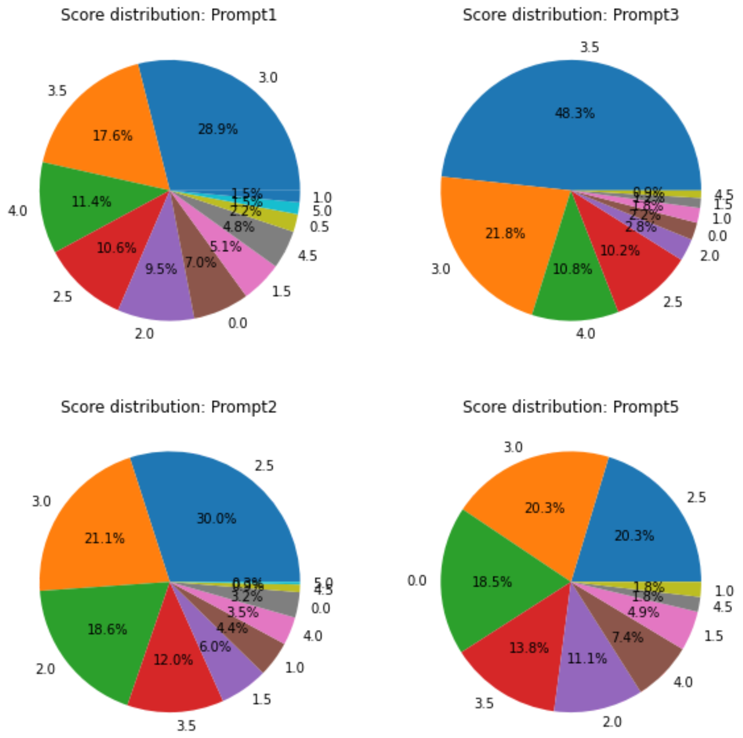

Fig 1: Score distribution is uneven for prompts

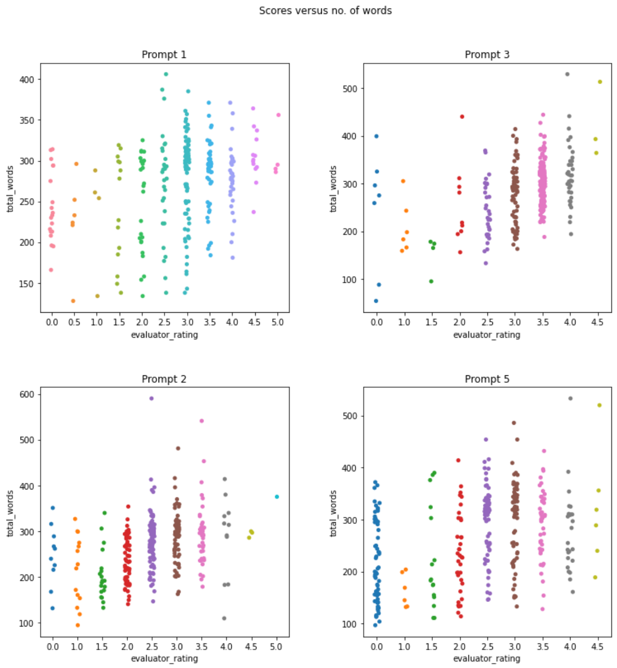

Fig 2: Trend of score versus number of words

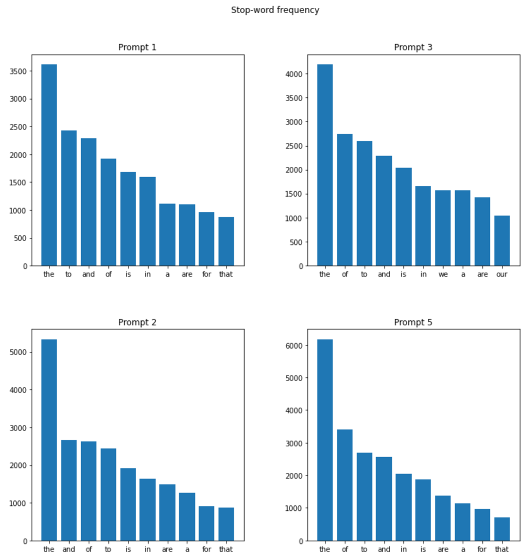

Fig 3: Stop Word frequency for each prompt

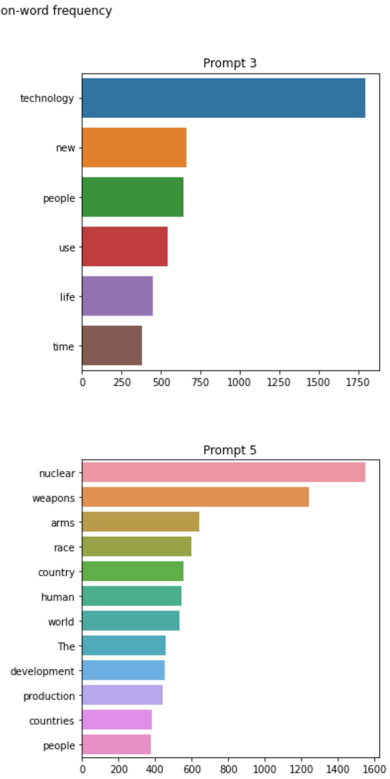

Fig 4: Bigram analysis

Data Cleaning

The following steps were done in data cleaning:

1. Lowercasing
2. Removing punctuations and digits
3. Correcting spellings
4. Removing StopWords
5. Lemmatisation

Feature Engineering

This step focuses on converting text sentences to vectors. The essays were tokenised (separately for each prompt) and then converted to a vector of numeric sequences (with padding)

DL model

4 separate models were developed (one model for each prompt). The model has the following properties:

1. **Glove embedding layer(100):** A non-trainable layer that is used for yielding word-vectors with lesser dimensions
2. **LSTM layer (128):** In order to reap the benefits of the sequence-based structure of an essay, LSTM is implemented
3. **Dense (32):** To learn the featurettes of scoring
4. **Dense (1):** The output layer which points to the evaluator rating

Specifications

The model is trained with MSE loss as it best-suits regressive models. The model is also trained with an Adam optimiser. The number of epochs is fine tuned from 100 to 50 to 20, and the number of folds is fine tuned from 5 to 3 for K-fold cross validation.

GitHub

https://github.com/kodooraKILLER/the_NLP_multiverse