



Anmol Gupta

Annie Hwang



Automatic Grading for Essays



Paul Lisker

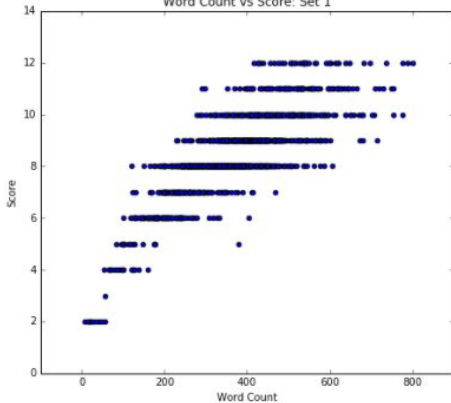
Kevin Loughlin

Data Exploration

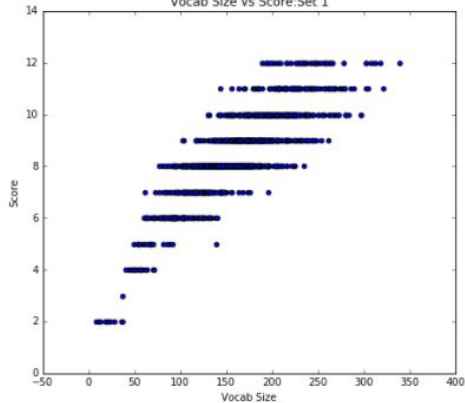
- 8 Essay Sets
- 12,976 Training Essays
- 4,218 Testing Essays
- Each Essay Set has different topic and scoring system
- ISO-8859-1 text encoding to handle special characters
- No missing data (yay!)
- Explored possible features of interest, such as words counts and vocabulary size vs. the essay scores for each of the datasets

Comparing Essay Sets

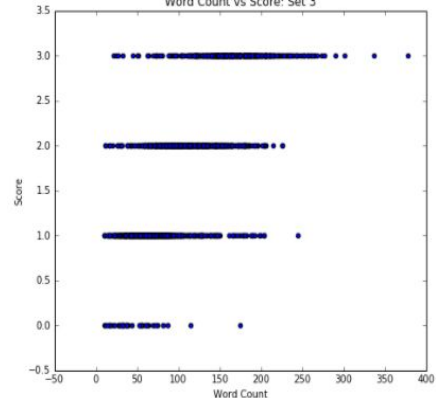
Word Count vs Score: Set 1



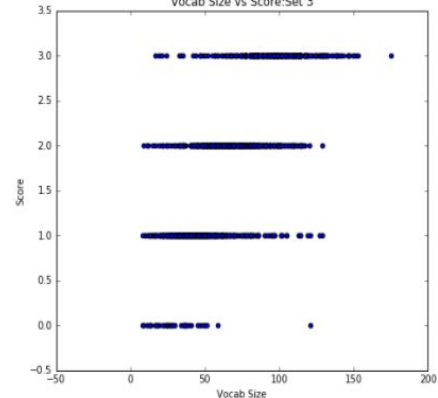
Vocab Size vs Score: Set 1



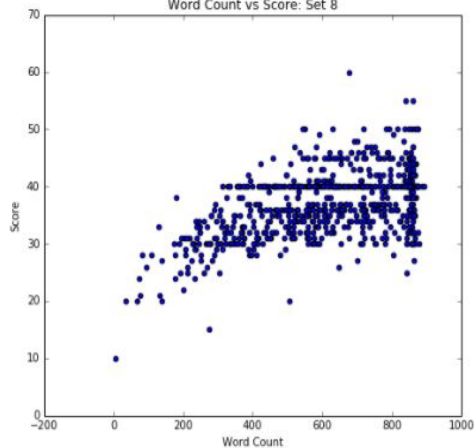
Word Count vs Score: Set 3



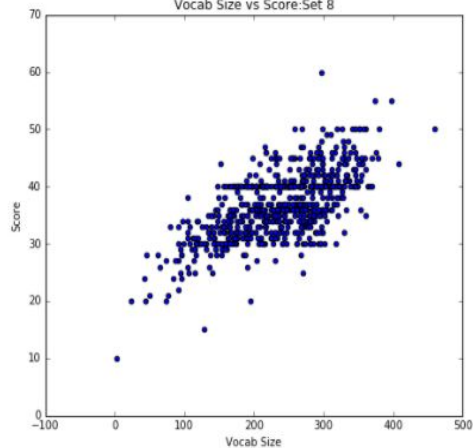
Vocab Size vs Score: Set 3



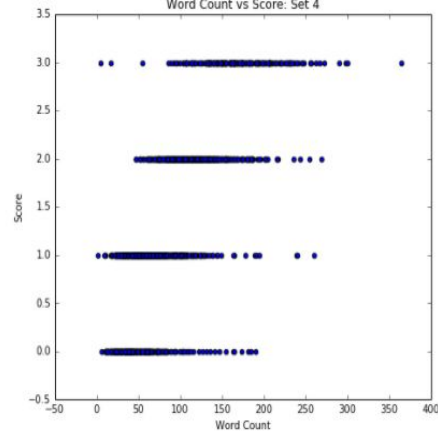
Word Count vs Score: Set 8



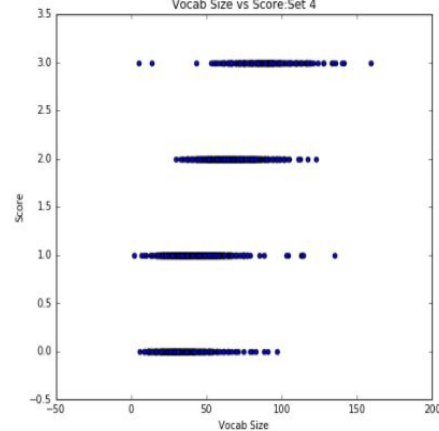
Vocab Size vs Score: Set 8



Word Count vs Score: Set 4



Vocab Size vs Score: Set 4



Baseline Models

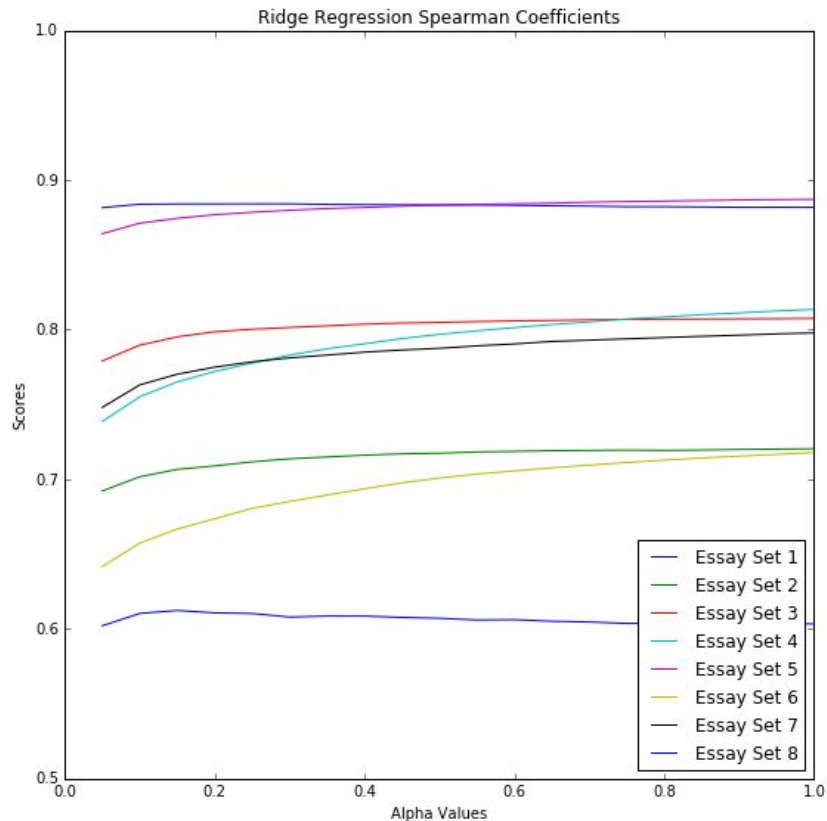
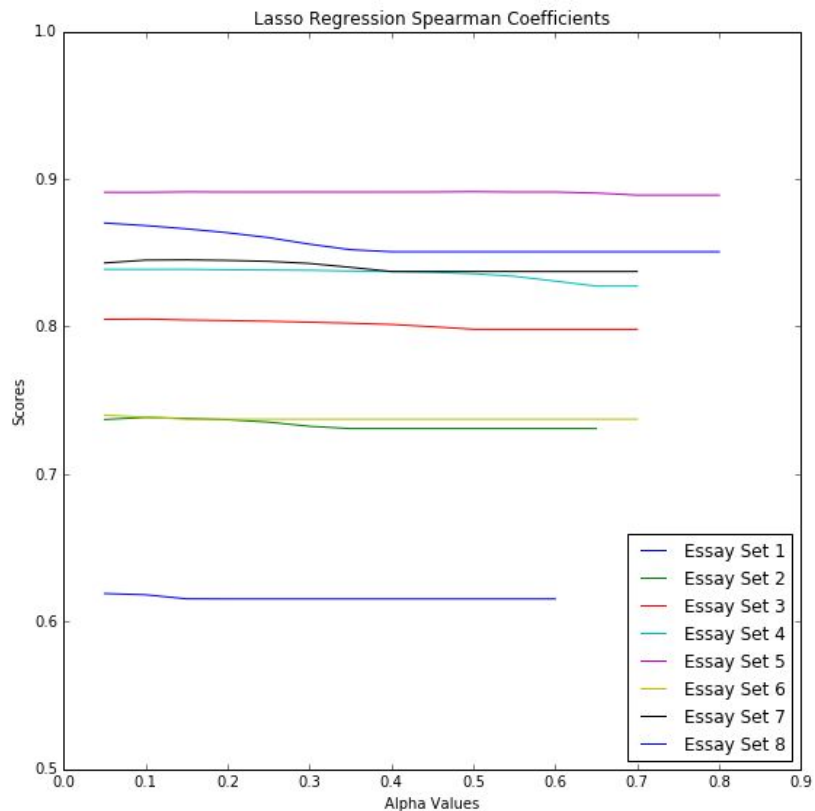
- Standardized the scores of the 8 essay sets, so that the mean = 0 and std = 1 within the sets.
- Used unigrams based on the TFIDF counts in the training set
- Calculated the Spearman Coefficient of our predicted scores.
- Baseline results are listed below

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Spearman	0.483218	0.374053	0.314986	0.316278	0.405456	0.151559	0.402221	0.394887
p-value	8.613e-36	2.311e-21	1.506e-14	4.429e-15	3.473e-25	0.000194	1.405e-18	4.073e-10

What We Learned from Baseline Model

- Generally, for the Spearman Coefficient,
 - .00-.19 “very weak”
 - .20-.39 “weak”
 - .40-.59 “moderate”
 - .60-.79 “strong”
 - .80-1.0 “very strong”
- All p-values below 0.05, meaning that there is a monotonic correlation between all of the essay sets
- Hence, generally moderate (but significant!) scores for baseline model
- Need to try other features! (Perplexity, % P.O.S., # Sentences, % Misspellings, # Words, # Unique Words)
- Need to try regularization via Ridge and Lasso Models (alpha ranging from 0.05 \rightarrow 1.0)

Visualizations



Results

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Average
Alpha Values	0.3	1.0	1.0	1.0	1.0	1.0	1.0	0.15	
Max Ridge	0.884028	0.720432	0.807483	0.813519	0.887043	0.717841	0.797747	0.612373	0.780058
Alpha Values	0.05	0.1	0.1	0.15	0.50	0.5	0.15	0.05	
Max Lasso	0.870108	0.738401	0.805037	0.838745	0.891339	0.739961	0.84513	0.619065	0.793473
LinReg Scores	0.780988	0.572063	0.017910	0.590328	0.719688	0.460429	0.611579	0.509009	0.532749

Interpretation

- Lasso Regression, on average, gives a higher score than Ridge.
- The p-values (not displayed) for all of the scores are significantly lower than 0.05, meaning that each of the essay sets have a monotonic correlation!
- The Spearman scores are all “strong” or “very strong,” demonstrating the success of adding the predictors.
- The Ridge Regression Scores were associated with much larger alpha values on average, meaning regularization really helps!
- Our new correlations are ~2x those that we obtained with our baseline model

Future Work

- NLTK perplexity functionality (bigram and trigram for perplexity)
- Overcoming computational power and time limits
- Using Parse Trees as features
- Taking the essay prompt into account

