Better Reading Levels through Machine Learning

Adam Gall: aag2113@stanford.edu; CS229

Dec 12, 2014

Introduction

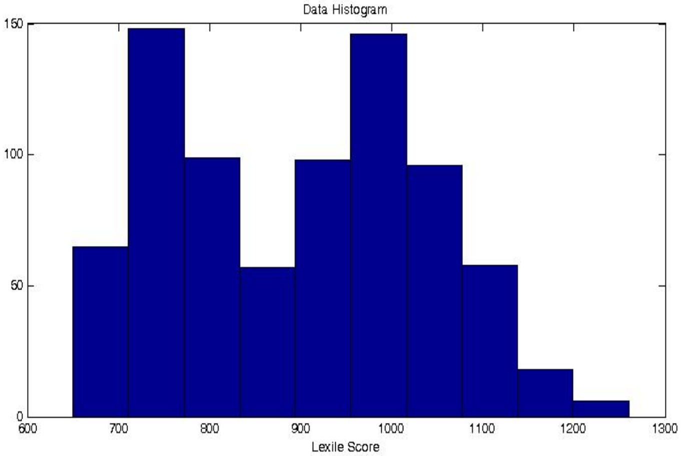
Measuring the reading difficulty of a particular text is a common and salient problem in the educational world, particularly with respect to new/struggling readers. While “common sense” measures exist for canonical texts, assigning an appropriate reading level metric to new resources remains challenging. Current systems have been widely criticized for misrepresenting the difficulty of texts which causes frustration for students and educators alike. Currently the most popular metric is the Lexile Reading Measure which is both proprietary and expensive. I aim to use machine learning to reproduce the results of the Lexile Measure and hopefully improve it with the insights gained through this exercise.

Dataset

One major hurdle in constructing a data set for this project is obtaining texts that have been rated by Lexile. This is critical as the cost to have a new text rated is prohibitive. At the outset, I had a set of 800 texts that were rated by Lexile. These texts are approximately 1500 words in length and have lexile ranges from 650 to 1260.

I intended to pad this set with additional texts from project Gutenberg (an online collection of public domain books and articles) however, I learned quickly through analysis that there are insufficient texts available through project Gutenberg that have Lexile ratings to sufficiently pad my dataset. Thus deeming this source unhelpful. Furthermore, I was unable to find another source of large numbers of Lexile rated texts and thus was left with my original 800 texts.

While my dataset is fairly robust, it does have several issues. First, Lexile ranges from 200L-1600L but the vast majority of my texts range from 650L-1260L and clusters around 750 and 1000 so it is possible that I will overfit this segment of the scale. Furthermore, my texts are predominantly biographical encyclopedia articles from a small number of authors. These are also potential sources of bias.



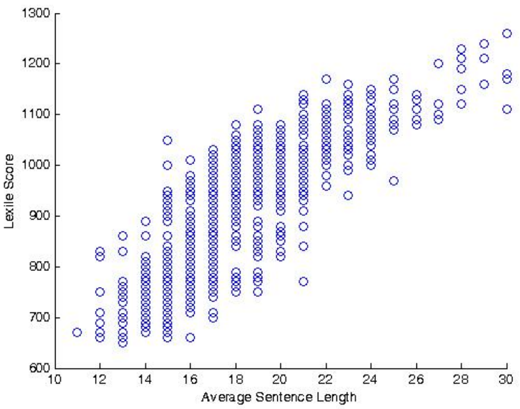
Features and Preprocessing

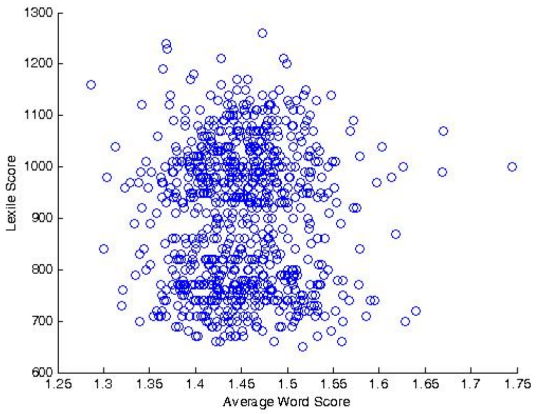
I have initially focused on four features: sentence length, paragraph length, word length and difficulty of vocabulary. Preprocessing for these features is relatively simple with the most effort going into constructing a vocabulary list. I have chosen to use Python's nltk module for natural language processing and have used it to assign part of speech and also to stem the words in my vocabulary list. Furthermore, I removed articles, conjugations and prepositions as they do not weigh heavily in any of the open reading measure standards and are unlikely to provide much information. As one may expect there are a large number of words that appear very infrequently in the data. I have ultimately decided to include these as they may be difficult words that contribute significantly to the reading level.

In order to determine vocabulary difficulty I cross-referenced several standard state vocabulary lists by grade level focusing on "sight words" (words that cannot be "sounded out" and thus must be memorized). This left me with a list of 2076 words rated by grade level 0-12. Naturally, this is a small sample of the 25,000 words that are present in the data. However, reading comprehension relies heavily on understanding of vocabulary, thus any proxy that we may be able to find for understanding should be useful.

From the outset I expected a strong correlation between sentence length and Lexile score while I assumed that vocabulary difficulty would be a harder feature to incorporate, this assumption is supported by the data.

Finally, I created 7 discrete classes for the Lexile measures (6-12) to increase the universe of available algorithms. I also believe that reading level is itself an imprecise concept and would ultimately be better represented as a distribution rather than a concrete number. Furthermore, the Lexile scale roughly corresponds to grade level which is how it is most commonly used/referred to.





Models

I reserved 30% of my data for testing and utilized the remaining 70% to train.

Locally Weighted Linear Regression

Because the correlation between sentence length and Lexile score appeared strong in my initial plots I had high hopes that adding the right feature would yield strong results for this model. It follows that the most challenging aspect of this model was determining which features were likely to provide additional information. I tried numerous combinations of my features and ultimately found the strongest results came from using average sentence length as the only feature.

K-Means Clustering

To implement this algorithm I started by separating the data into 7 clusters (one for each discrete lexile score represented in the data). Once this was complete I created a probability distribution from the actual Lexile scores of the documents in each cluster. Finally, I assigned new documents to the most likely Lexile score for its given cluster. I attempted this approach using various feature vectors from my full feature vector, to just words to just words for which I had a grade level and was unable to achieve results that much better than a coin flip.

Below is the probability density generated by using K-means on the word vectors generated from the raw texts. Clearly, clustering in this fashion fails to generate definitive classifications. However, it is interesting to note that the densities roughly center around 700 and 100L as we would expect from the word-score scatter seen above.

Class/Lexile	6	7	8	9	10	11	12
1	.0128	.141	.1282	.3077	**.3462**	.0641	0
2	0	.1556	.1333	.2444	**.3556**	.1111	0
3	0	.1739	.1304	**.4783**	.1739	.0435	0
4	.0127	**.2866**	.1656	.2420	.2166	.0764	0
5	.390	.2338	.2208	**.2597**	.2208	.0260	0
6	0	.1290	.0645	.0968	**.3871**	.1935	.1290
7	.1328	**.4844**	.1797	.1172	.0547	.0313	0

Support Vector Machine

SVM provided the strongest results of any algorithm that I implemented. However, even here we hardly performed better than the flip of a coin. One interesting observation that arose from SVM testing was that the algorithm performed similarly when given only the word frequency features. This indicates that there is indeed information present there that I have been unable to effectively describe in feature generation and furthermore gives me hope in my continued efforts.

Results

Below are the best results that I achieved with each model after exhaustively testing different combinations of features. While my exact classifications are inconclusive it is worth noting that both LWLR and SVM performed very well at classifying each text within one Lexile level. Clearly, Lexile is using a feature that I have been unable to discover but we are able to recommend texts that are approximately appropriate within one grade level with near-certainty.

Model	Training Error	Within one class error
LWLR	.5476	.013
SVM	.4563	.0159
K-Means	.6508	.1468

Future

I am very passionate about this subject and will continue to improve upon the work that I have done in an attempt to create a better system for predicting reading difficulty. One thing that was unable to sufficiently explore within the scope of this study is the effect that parts of speech, conjugations, proper nouns, numbers and arguably objectionable words lend to reading level. These features should be relatively simple to analyze using modern language processing tools. However, I expect that the impact of such features will be difficult to quantify directly which is why I avoided them for this project.

Finally, I believe strongly that the best measure of reading level will incorporate textual analysis to classify things like concept and tone that have traditionally eluded quantification. This was outside of the scope of this study, yet I do not believe that we will have a truly effective system that works for both students and educators until such features are incorporated.

References

"Dolch word list." Wikipedia, n.d. Web. 15 Nov. 2014.

Fry Edward. Dr. Fry's Spelling Book Levels 1-6. Teacher Created Resources, 2000. Print

"Sight Word Lists." Student Resources. Tarpey Elementary School, n.d. Web. 15 Nov. 2014.

"Sight Words: Grades K, 1, 2 & 3." Pawnee CUSD 11. Simplified Online Communication System, n.d. Web. 15 Nov. 2014.

U.S. Department of Education, Office of Career, Technical, and Adult Education (OCTAE). "Dolch Basic Sight Word List." Literacy Information and Communication System (LINCS), n.d. Web. 15 Nov. 2014.

"WISD High Frequency Word Lists by Grade Level." Waxahachie Independent School District, n.d. Web. 15 Nov. 2014.

“Assessing the Lexile Framework: Results of a Panel Meeting.” National Center for Education Statistics, Aug 2002.

“How Lexiles Harm Students.” Mike Mullen, October 22, 2012
http://mikemullin.blogspot.com/2012/10/how-lexiles-harm-students.html