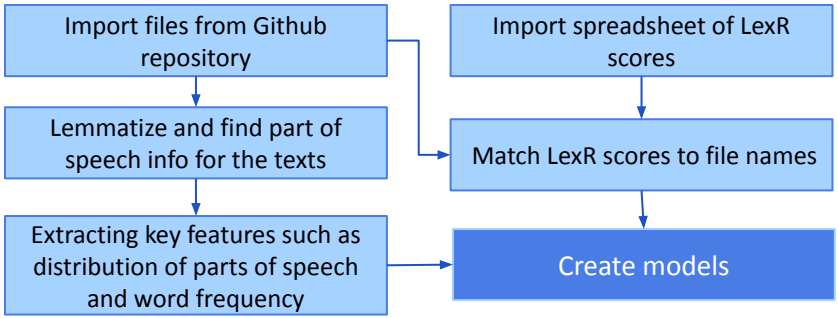
Calculating Latin Readability Scores Using Linear Regression

Amanda Lin, Surya Kolluri, Dylan Sheehan Faculty Mentor: Prof. Bret Mulligan MehtA+ Instructors: Mohammad Sharafat, Haripriya Mehta

Motivation

- There currently isn't any method to automatically categorize the complexity of Latin texts while there are models for the majority of modern languages.
- How can machine learning be used to create a model that will score the difficulty of Latin texts?
- How can our model classify texts based on features such as LexR scores?

Methods



Experiments

- Models used: Linear regression, ridge regression, ElasticNet and ElasticNetCV
- Parameters changed: Number of iterations (epochs), Alpha values, Ratio of L1 to L2 regularization

Results

Model Type	Linear Regression	Ridge Regression	ElasticNet	ElasticNetCV
Accuracy Score	0.728	0.728	0.725	0.667

Conclusion & Future Work

- Models all produced similar results
- Models can be improved for higher accuracy
- Different types of models can be tried
- Include more features of texts

References

Sarah E. Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language* 23(1):89–106, 2009.

Sarah E. Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 523–530, USA, 2005. Association for Computational Linguistics.

Kevyn Collins-Thompson and James P. Callan. A language modeling approach to predicting reading difficulty. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* , pages 193–200, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

HAVERFORD COLLEGE

A+