Predicting Success of Literary Fiction
Project Proposal

Joy Zhong and Will Zhou

I. PROBLEM

We aim to build a model that will predict the success of literary works. While researchers have studied similar topics such as readability in academic articles [1] and quantitative assessment of English fluency in text [2], the contrast between good and excellent writing in fiction is a relatively novel task. In 2013 Ashok et al. built an SVM-based model which predicted the success of literary works with up to 84% accuracy [3]. We plan to approach the same problem using random forests because of the ease of interpretation that decision trees give in regards to features.

One motivation for predicting the success of novels comes from the needs of publishers to quickly distinguish potentially popular books from the thousands of manuscripts they receive. Additionally, we hope that by using random forests we will be able to isolate certain features of text such as vocabulary, elements of lexical cohesion, or syntax that correlate with successful or popular writing. Although there exists a wealth of literature on good writing based on qualitative observations, there are fewer studies based on quantitative analyses, perhaps because of the complexity of the problem and the numerous factors that play into the success of a novel. In this study we hope to isolate at least a few of these factors to reveal new insight into what constitutes "good" and successful writing.

II. DATA

We plan on using data from Project Gutenberg[1]. Project Gutenberg offers more than 46,000 e-books that can be freely downloaded. In addition they include for each book the author, time period, language, and genre or sub-genres. In order to control for writing style influenced by era, we will use only books written between 1800 and 1950. Furthermore, although we will initially use books from every genre for our experiments, we may eventually extend our project to be genre-specific, as previous studies have done so relatively successfully [4].

To measure the success of the novels we will use download count overall. Since this count is affected by the date which the book was made available on Project Gutenberg we will also perform experiments using the number of downloads over the past 30 days. Additionally, following Ashok et al. we will limit the number of books by the same author in our data set to two, in order to prevent our model from learning based on author style rather than linguistic content [3].

[1] *http://www.gutenberg.org/*

III. METHOD

A. General Approach

The data will be partitioned into three subsets: training, development, and test, with each subset containing 60%, 20%, and 20% respectively. Plain-text input will be trained on to predict a download count for each literary work.

B. Random Forests

Although much of the literature within the text classification space employs SVM's and SVR's as its preferred machine-learning model, we are considering the use of random forests. The motivation for this is two-fold. Firstly, decision trees are generally very easy to interpret, and the bagging and randomized feature selection methods used to build random forests greatly mitigate the propensity of decision trees to overfit to training data. Secondly, random forests appear to perform well when compared to other classifiers such as neural networks or SVM's [5].

As a comparison, we plan to implement the SVM model by Ashok et al. with the same parameters, to gain insight on how random forests perform versus SVM's in this particular text classification problem [3].

C. Feature Selection

As baseline textual features, we will consider unigrams and bigrams, part-of-speech tagging, syntactic features such as distribution of pronouns or "Wh-" words, structural features such as sentence length in terms of words or characters, and sentiment. Genre is one non-textual feature that will also be incorporated.

Previous literature in similar areas have found distribution of constituents, distribution in connotation [3] to be useful, and in terms of readability, difficulty in vocabulary, syntax, discourse relations, and elements of lexical cohesion have demonstrated importance [1]. We hope to implement these features as well and compare performance of our base model to a model with the entire aggregate.

IV. MILESTONE GOALS

By January 27, we will finish scraping all the data from Project Gutenberg and understand the different variations of random forests. By February 3, we will have a working implementation of random forests with several baseline features. Before the milestone, we will finish the full model for our project. Once we have reached the milestone, we will begin analysis of the model under different parameters and run different comparisons.

REFERENCES

[1] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

[2] Nenkova, A., Chae, J., Louis, A., & Pitler, E., Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text, Empirical Methods in Natural Language Generation: Data Oriented Methods and Empirical Evaluation, 2010.

[3] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP*.

[4] Annie Louis and Ani Nenkova. 2013. What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of ACL*.

[5] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R news, 2(3), 18-22.