Machine Learning Literature using Textual Features

van Cranenburgh, Andreas

Published in:
Tiny Transactions on Computer Science

Publication date:
2016

Machine Learning Literature using Textual Features

Andreas van Cranenburgh
Huygens ING & Institute for Logic, Language and Computation

ABSTRACT

Literature is hard to define. The value-judgment definition holds that literature is a highly valued kind of writing [2, p. 9], but how arbitrary or predictable are such judgments? Moreover, some believe that critics and publishers wield more influence than the text itself [1]. We investigate these questions with a computational model of literature trained on texts.

As part of The Riddle of Literary Quality (`http://literaryquality.huygens.knaw.nl`), an online survey (14k respondents) was conducted among the general public to collect judgments on 401 recent, bestselling Dutch novels. Given a list of author-title pairs, respondents rated novels they had read on a 7-point scale from definitely not to highly literary.

We consider the regression task of predicting the mean rating of each novel using features extracted from its text. We train a linear support vector regression model on frequencies of bigrams and syntactic features. The syntactic features consist of tree fragments mined from trees obtained by automatically parsing the novels.

Our predictive model explains 57.5 % of the variance in literary ratings, with a root mean squared error of 0.65 on a scale of 0–7 (evaluation based on 5-fold cross-validation with the 401 novels). This is in line with pilot experiments with a subset of the novels and only bigrams [3]. Although the bigrams form a simple, strong baseline, the syntactic features are more interpretable.

We conclude that perceptions of literary ratings can be explained to a large extent from the text itself: there is an intrinsic *literariness* to literary texts.

BODY

Why are some novels considered to be literature? A predictive model of literary judgments shows that textual features are important.

REFERENCES

[1] P. Bourdieu. *The rules of art: Genesis and structure of the literary field.* Stanford University Press, 1996.

[2] T. Eagleton. *Literary Theory: an Introduction.* University of Minnesota Press, Minneapolis, 2008.

[3] A. van Cranenburgh and C. Koolen. Identifying Literary Texts with Bigrams. In *Proc. of workshop Computational Linguistics for Literature*, pages 58–67, 2015. `http://aclweb.org/anthology/W15-0707`

Volume 4 of Tiny Transactions on Computer Science