1. Introduction

Text classification is the process of categorizing text based on features found in the text itself. Features may pertain to form, such as sentence length and number of syllables, or they may pertain to subject matter.

The goal of this project was to correctly classify poems in two ways. First, as rhyming or not, and then as following a meter or not. The meter chosen was the iambic pentameter, which has five sets of unstressed and stressed syllables, i.e. da DUM da DUM da DUM da DUM da DUM.

2. Methodology

- Scour the web for the poetry
- Extract poetry using a python module called Beautiful Soup
- Generate features with the CMU pronunciation dictionary and a python module for calculating Levenshtein distance
- Run features in WEKA to train a classifier

3. Results

Classification: use of iambic pentameter

- Average Number of Stressed Syllables Per Line
- Average Number of Syllables Per Line
- Average Levenshtein Distance from the Ideal

Classification: use of rhyme

- Percentage of rhyming lines
- Number of lines in poem

4. Conclusion

The initial results were promising. Adherence to a poetic meter and use of rhyme are fairly easy to detect. A large amount of time was dedicated to text mining, feature extraction and establishing a classification work-flow.

Future work will involve improving accuracy of rhyme and meter detection in order to classify poems into multiple distinct rhyming and meter classes.

5. References


6. Contact

Bryan Paget
bdjpaget@gmail.com

Dr. Diana Inkpen
Diana.Inkpen@uottawa.ca

Dr. Chris Tanasescu
margento.official@gmail.com