Machine learning in poetry classification

Bryan Paget
Supervisors: Dr. Diana Inkpen, Dr. Chris Tanasescu

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

Beautiful Soup

Generate features with the CMU pronunciation dictionary and a python module for calculating Levenshtein distance

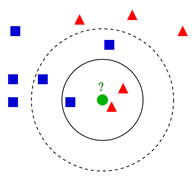
The CMU Pronouncing Dictionary

Shall I compare thee to a summer's day?

SHAE**1**L AY**1** KAH**0**MPEH**1**R
DHIY**1** TUW**1** AH**0**
SAH**1**MER**0**EH**1**S DEY**1**?

11011101011 01011101011 01010101011 0101010101
0101010101 0101010101 0101010101 0101010101

Run features in WEKA to train a classifier



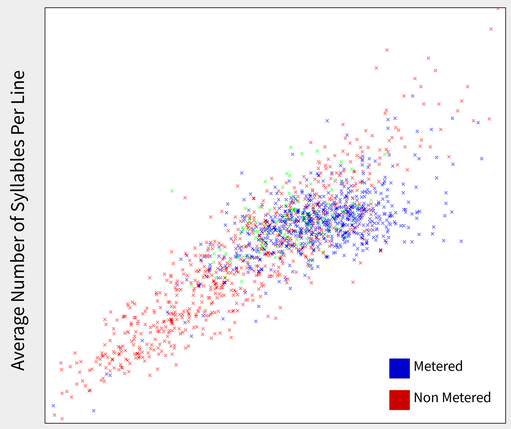
Build a user application

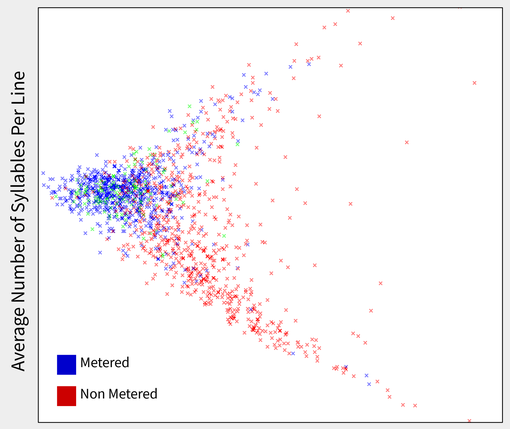
3. Results

Classification: use of iambic pentameter

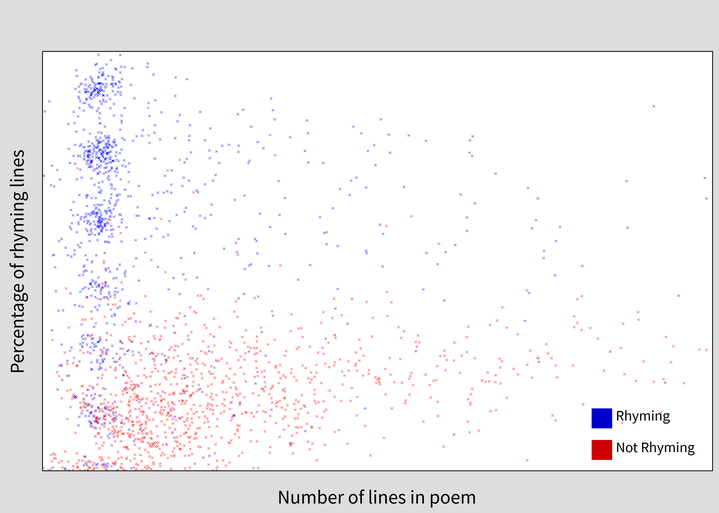
Classification Matrix		
Classified as:		
Metered	Non Metered	Actual Class:
737	78	Metered
106	599	Non Metered

~88% accuracy with J48





Classification: use of rhyme



Classification Matrix		
Classified as:		
Rhyming	Not rhyming	Actual Class:
932	170	Rhyming
133	1145	Not rhyming

~88% with OneR

Example data:

Sonnet 18

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date.
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed;
And every fair from fair sometime declines,
By chance, or nature's changing course, untrimmed;
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wand'rest in his shade,
When in eternal lines to Time thou grow'st.
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

-- William Shakespeare

11011001011
111100110
1111010101
01011111101
1211010101
010001010
01001011201
110101101
1101010111
1101000011
11111001
100101011
1111010101
1110001101

Sonnet 18	
Characters	639
Words	114
Lines	14
Stanzas	4
Syllables	126
Lines that rhyme	~ 0.43
Syllables Per Line:	
Average	~ 9.69
Max	11
Min	8
Median	10
Variance	~ 0.83
Standard Deviation	~0.91
Stressed Syllables Per Line:	
Average	~6.15
Max	8
Min	3
Median	6
Variance	~ 1.5
Standard deviation	~ 1.2
Levenshtein distance	
Average	~ 2.92
Max	5
Min	2

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

Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco.

J. G. Cleary and L. E. Trigg, “K*: An Instance-based Learner Using an Entropic Distance Measure,” Mach. Learn. Work. Then Conf., vol. 5, pp. 1–14, 1995.

R. Bilisoly, “Quantifying Prosodic Variability in Middle English Alliterative Poetry,” pp. 1230–1241, 2007.

6. Contact

Bryan Paget
bdjpaget@gmail.com

Dr. Diana Inkpen
Diana.Inkpen@uottawa.ca

Dr. Chris Tanasescu
margento.official@gmail.com