Marla Seth
ibc346
Dr. Zanella
IS 6713 – Spring 2021

What do you want to do?

Poetry is among the oldest genres of literacy and is often the earliest record we have of most cultures. It is a form of expression which can greatly enhance the social-emotional learning of students. However, poetry units in the secondary classroom are shrinking or even disappearing altogether. The most common justifications for shortening or even skipping poetry units completely are that teachers often feel ill-equipped to teach units on poetry, teachers feel somewhat apathetic to the genre, teachers feel students are reluctant to engage with the genre, and teachers feel there is insufficient time for poetry units when so much emphasis is placed on standardized test preparation (Young, 2016).

Because teachers feel there is insufficient time to teach poetry, and therefore feel the necessity to shorten poetry units, choosing the appropriate poems for the unit is essential. Chosen poems must accessible and engaging to students. Choosing ballads specifically could be the solution to a lack of student engagement. Because ballads are narrative poems that focus on a specific story told through action and dialogue, students may find it easier to connect with the poem. The sing-song nature of the verse can be evocative of their own music. Additionally, the frequent repetition of words and phrases throughout stanzas could make the poem feel more familiar and accessible to students. This project will attempt to write an algorithm that can analyze the text of an online poem and classify it as either a ballad or not a ballad.

Why should we care?

Many poems found online are misclassified as ballads. For example, Edgar Allen Poe's "Annabel Lee" is very frequently classified online as a ballad because Poe himself referred to the poem as a ballad. However, "Annabel Lee" is missing a key characteristic of ballads; it lacks dialogue. Because time is such a precious commodity for teachers, it would be beneficial to have an automated "pre-screening" process for poem selection. If an algorithm can correctly identify a poem as a ballad, the teacher can then focus on reading and picking the best ballad for the class without first wasting time classifying the poems.

What data will you use:

All data for this project will be scraped from the Poetry Foundation's website (https://www.poetryfoundation.org). Three specific poems will be used to establish criteria for a ballad. Those poems are Sir Walter Scott's "Lady of the Lake: Boat Song", Lord Alfred Tennyson's "Lady of Shalott (1842) ", and Edgar Allen Poe's "The Raven."

How will you determine success?

The algorithm will be tested against eighteen poems that the Poetry Foundation has classified and misclassified in a filtered list as ballads suitable for children. The measure of success of will be the algorithms accuracy as compared to a manually labeled list of the same poems.

How will the data be processed and analyzed?

To identify a poem as a ballad, the algorithm will focus on the following two characteristics of ballads: use of dialogue and high repetition of words and phrases. Specific rhyming schemes will not be analyzed for this analysis and poems that meet the other two criteria without specifically adhering to the traditional ballad rhyming scheme will still be considered as ballads. The three poems to be used for criteria will be scraped from their respective website, cleaned, and prepared for analysis. Quotation mark usage will be counted to check for usage of dialogue. High-frequency, irrelevant words will be removed, and the frequency of other words will be counted. A percentage of the most repeated words as compared to the total number of words (excluding high-frequency, irrelevant words) will be calculated. The percentages of repeated words for the three poems will then be compared to establish a minimum amount of repetition necessary to be considered a ballad. Once the criteria are established, the text from the eighteen poems will be analyzed against those criteria and classified as either a ballad or not a ballad. Finally, the classifications will be compared against the manually labeled list.

How much will it cost and how long will it take?

The Poetry Foundation's website is free to access, and the free version of Jupyter Notebook will be used. Therefore, there will be no monetary cost to this project. However, there will be a cost of time. Given my novice experience with python, I estimate the project will take me between twenty-five and thirty hours. A breakdown of the time estimate is below:

Manually labeling the set of 18 test poems – 2 hours
Writing and troubleshooting code using only one of three criteria poems: 8 hours
Expanding code to include all three criteria poems: 1 hour
Comparing results from the three poems and developing criteria: 1 hour
Writing and troubleshooting code using one poem of eighteen test poems: 6 hours
Expanding code to include all eighteen test poems: 2 hours
Comparing code results the manually labeled list: 1 hour
Creating necessary visualizations, figures, and/or tables: 4 hours
Writing the final report: 2 hours
Total: 27 hours

WORK CITED:

Young, M.A. (2016). High School English Teachers' Experiences with Poetry Pedagogy. (Unpublished doctoral thesis.) College of Professional Studies, Northeastern University, Boston, Massachusetts. Retrieved from https://repository.library.northeastern.edu/files/neu:cj82n331g/fulltext.pdf