Marla Seth
ibc346
Dr. Zanella
IS 6713 – Spring 2021

What's a Ballad? The Development of an Algorithm to Predict Ballad Poetry

Introduction

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 be 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.

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.

The Data

Text for all poems was scraped from the Poetry Foundation website. Three ballads and three non-ballads were chosen for analysis to identify quantitative criteria that will be used to classify a ballad. The ballads chosen were 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." The three non-ballads chosen were Robert Frost's "After Apple Picking", Jill Alexander Essbaum's "Parting Song", and Terri Kirby Erickson's "Fund Drive." After criteria were chosen and the predictive algorithm written, it was tested against a set of eighteen poems labeled as ballads suitable for children. Because the Poetry Foundation classifies any poem with a sing-song nature as a ballad, regardless of narrative nature, action, and dialogue, many of the poems in the set would not be classified as ballads by English educators. The set was manually annotated as "ballad" or "not a ballad," and the annotations were used to calculate the accuracy of the algorithm.

The Method

Data was obtained and processed using python through Jupyter Notebook. The code can be found on github. The full text of all poems was scraped and saved to one of two csv databases. The first database contains the initial six poems used to decide upon the criteria to classify a ballad. The second dataset contains the eighteen poems which were used to test the algorithm. All poems were given gold standard labels of "ballad" or "not a ballad." Because the datasets are small, only 24 poems total, a machine learning approach was not feasible. Instead, criteria were identified manually. When manually annotating the poems with gold-labels, a poem was considered to be a ballad if it was a narrative poem, contained dialogue, and had repetition of words or phrases that could be considered a refrain.

The six poems used to develop algorithm criteria were scraped and processed using the Requests and Beautiful Soup libraries. Prior to any other processing, the full text of each poem was parsed, and the number of quotation marks were counted. Next, the text was tokenized into individual words and common stopwords were removed. After initial analysis, it was discovered that the word "like" behaves similar to a stopword in poetry. It appears frequently, but carries very little meaning wherever it appears. Due to this discovery, the word "like" was added to the stopwords to be removed, and the poems were reanalyzed. Because the individual forms of words are essential to the patterns of the poetry, words were not stemmed or lemmatized for analysis. Unique words were compiled for each poem and the number of times the word appeared in the text was counted. The most commonly appearing words were identified.

The Analysis

The initial six poems revealed that all ballads contained quotation marks to identify dialogue. In sharp contrast, none of the other three poems had any quotation marks. As a result, the first criterion chosen for identifying a ballad was the poem must contain at least two quotation marks.

Figure 1

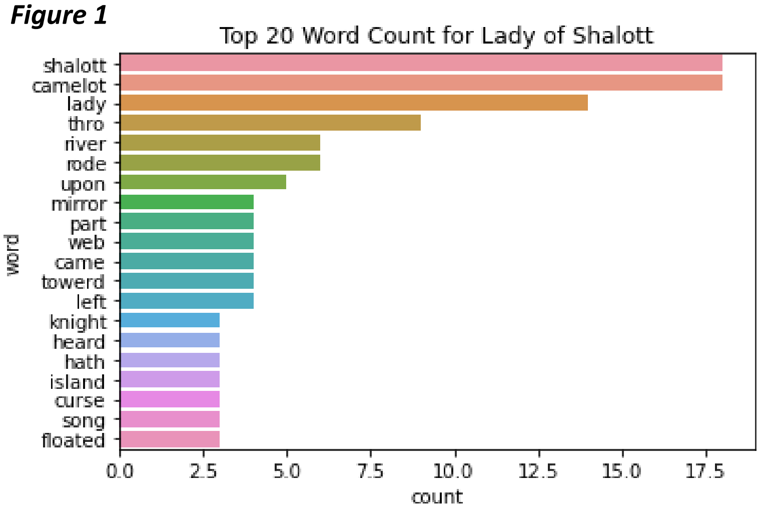


Figure 2

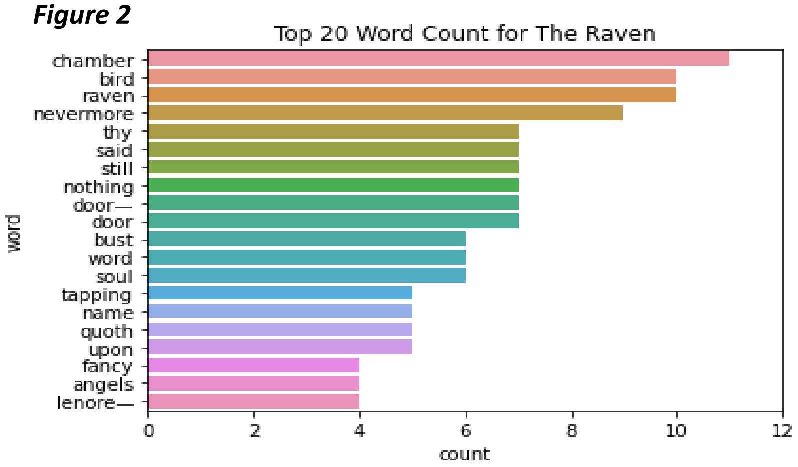
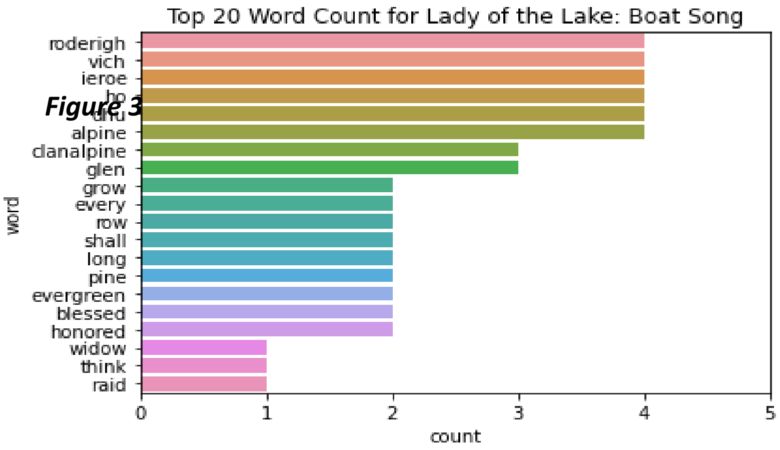


Figure 3



Next, the repetition of words and phrases was examined. To quantitatively identify a refrain, a ballad should have a set of words that appears much more frequently than other words within the text. Figure 1 shows the word count for the top 20 words in the ballad "Lady of Shallot." As the graph shows, there are three words, "shalott", "camelot", and "lady", that appear much more frequently in the poem. The number of occurrences for the following words drops significantly. Figure 2 shows the word count for the top 20 words in the ballad "The Raven." In this poem, there are four words that appear most frequently. The words are "chamber", "bird", "raven", and "nevermore". Again, the occurrences of other words starts to decline dramatically. Figure 3 shows the top 20 words in the ballad "Lady of the Lake: Boat Song." In this poem, there are six words that appear the most frequently. Those words are "roderigh", "vich", "ieroe", "ho", "dhu", and "alpine." Once more, the number of occurrences for the following words decline.

Figure 4 shows the top 20 words for the non-ballad poem “After Apple-Picking.” In this poem, there are two words, “sleep” and “apples”, that appear more frequently than other words. However, the rest of the words only appear once or twice in the poem. Figure 5 shows the top 20 words for the non-ballad poem “Parting Song.” In this poem, all notable words appeared only once. Figure 6 shows the top 20 words for the non-ballad poem “Fund Drive.” In this poem, the two words “take” and “father” appeared twice, but all other words appeared only once.

Figure 4

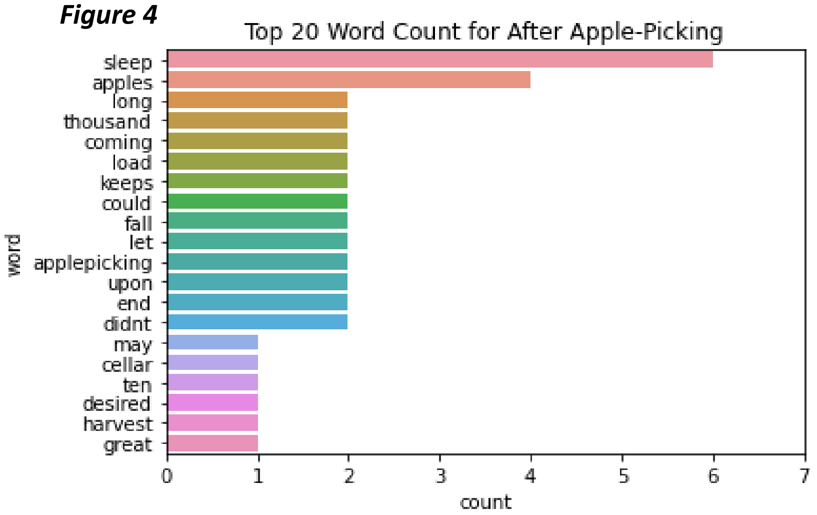


Figure 5

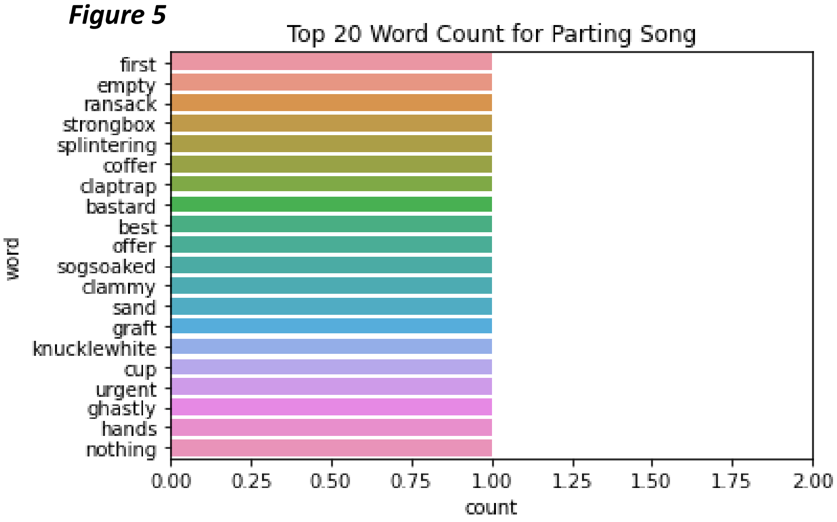
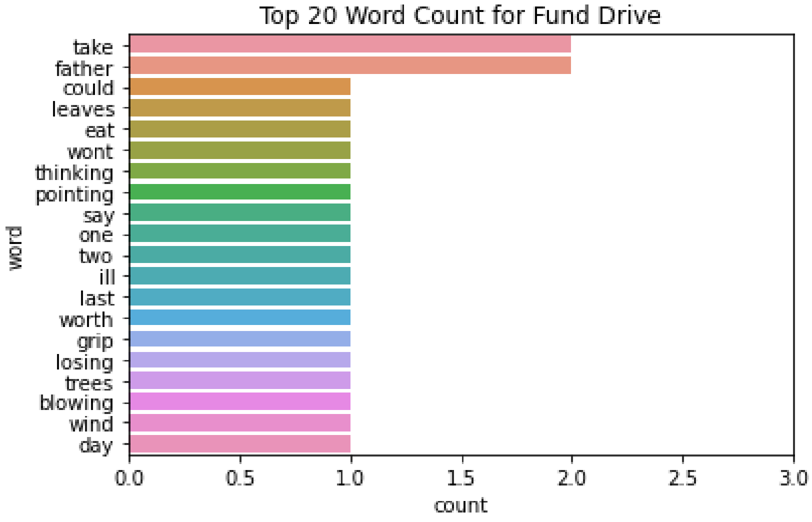


Figure 6



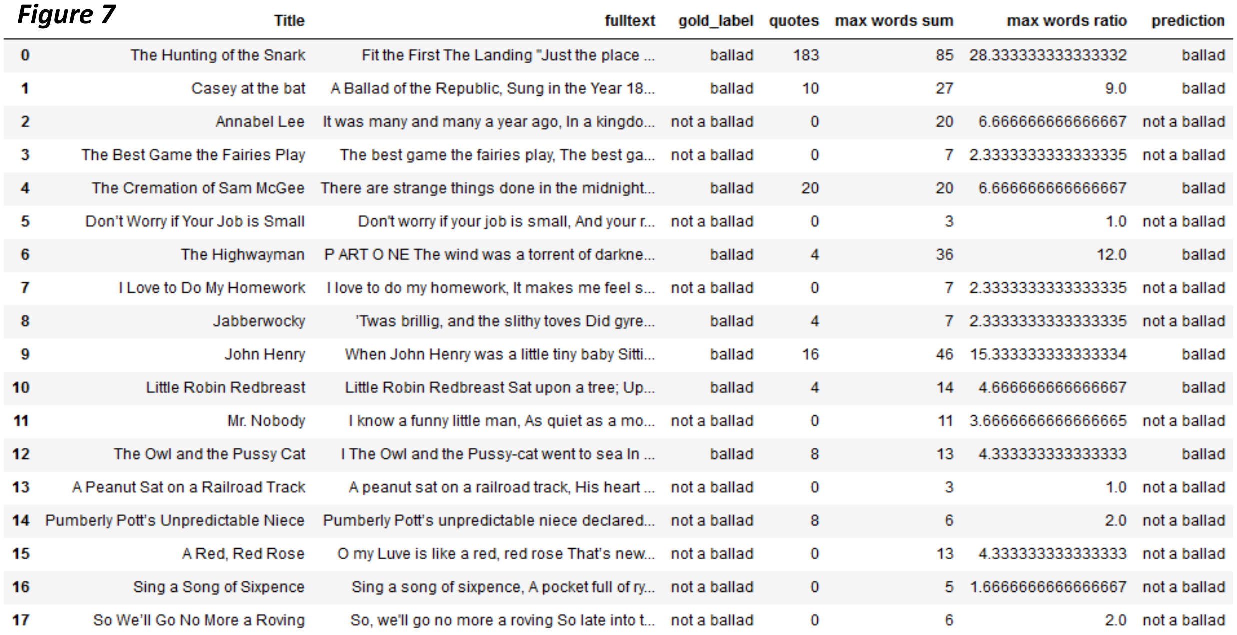
When examining all six graphs, a clear pattern emerged. Ballads do show more repetition in specific notable words than non-ballads. Because all the ballads had at least three words that were highly repeated and all the non-ballads had less than three words that were repeated, it was decided that the number of occurrences for the top three words in a poem was important. The average occurrence for the top three occurring words was calculated for each poem. In the ballad poems, the average repetition of the top three words was at least four times. Therefore, the second criterion chosen for the identification of a ballad was the average occurrence of the top three words must be greater than or equal to four.

The resulting algorithm processes the full text of the poem by first counting the number of quotation marks. Next, the text is tokenized and stopwords are removed. The number of occurrences of the remaining notable words are counted. The top three occurring words are identified. The average number of occurrences for the top three words is calculated. If the poem has at least two quotation marks and has an average occurrence for the top three words greater than or equal to four, it is labeled as a ballad. If it does not meet one of the criteria, it is labeled not a ballad.

The Results

The text of the eighteen test poems was analyzed and run through the algorithm to create predictions as to whether each poem was a ballad or not a ballad. The results are listed in the table in Figure 7. The sklearn package was used to calculate the accuracy, precision, recall, and F1 scores of the predictions as compared to the gold labels. The predictions from the algorithm matched the gold labels in all but one of the cases. The algorithm predicted the poem "Jabberwocky" would not be a ballad, when it had a gold label of being a ballad.

Figure 7



	Title	fulltext	gold_label	quotes	max words sum	max words ratio	prediction
0	The Hunting of the Snark	Fit the First The Landing "Just the place ...	ballad	183	85	28.333333333333332	ballad
1	Casey at the bat	A Ballad of the Republic, Sung in the Year 18...	ballad	10	27	9.0	ballad
2	Annabel Lee	It was many and many a year ago, In a kingdo...	not a ballad	0	20	6.666666666666667	not a ballad
3	The Best Game the Fairies Play	The best game the fairies play, The best ga...	not a ballad	0	7	2.3333333333333335	not a ballad
4	The Cremation of Sam McGee	There are strange things done in the midnight...	ballad	20	20	6.666666666666667	ballad
5	Don't Worry if Your Job is Small	Don't worry if your job is small, And your r...	not a ballad	0	3	1.0	not a ballad
6	The Highwayman	P ART O NE The wind was a torrent of darkne...	ballad	4	36	12.0	ballad
7	I Love to Do My Homework	I love to do my homework, It makes me feel s...	not a ballad	0	7	2.3333333333333335	not a ballad
8	Jabberwocky	'Twas brillig, and the slithy toves Did gyre...	ballad	4	7	2.3333333333333335	not a ballad
9	John Henry	When John Henry was a little tiny baby Sitti...	ballad	16	46	15.333333333333334	ballad
10	Little Robin Redbreast	Little Robin Redbreast Sat upon a tree; Up...	ballad	4	14	4.666666666666667	ballad
11	Mr. Nobody	I know a funny little man, As quiet as a mo...	not a ballad	0	11	3.6666666666666665	not a ballad
12	The Owl and the Pussy Cat	I The Owl and the Pussy-cat went to sea In ...	ballad	8	13	4.333333333333333	ballad
13	A Peanut Sat on a Railroad Track	A peanut sat on a railroad track, His heart ...	not a ballad	0	3	1.0	not a ballad
14	Pumberly Pott's Unpredictable Niece	Pumberly Pott's unpredictable niece declared...	not a ballad	8	6	2.0	not a ballad
15	A Red, Red Rose	O my Luve is like a red, red rose That's new...	not a ballad	0	13	4.333333333333333	not a ballad
16	Sing a Song of Sixpence	Sing a song of sixpence, A pocket full of ry...	not a ballad	0	5	1.6666666666666667	not a ballad
17	So We'll Go No More a Roving	So, we'll go no more a roving So late into t...	not a ballad	0	6	2.0	not a ballad

Because only one poem had a prediction that differed from the gold label, the algorithm has a very high accuracy score of 94.4%. The recall score, how many true ballads were correctly predicted, was the lowest at 87.5%. This is because "Jabberwocky" has a true gold label as a ballad, but the algorithm failed to predict it as one. The precision score, how many predicted ballads were indeed ballads, was the highest score at 100%. All true ballads were found by the algorithm. The F1 score, a composite score that balances precision and recall, was 93.3%. Overall, the algorithm did a great job of classifying ballads. However, due to the incredibly small sample size of poems, there are likely severe limitations to applying the algorithm to all poems.

The Costs

The Poetry Foundation website is free to access. Jupyter Notebook is also software that is free to download and use. Therefore, there was no monetary cost to this project. The only real cost was time invested. From start to finish, the project took approximately 30 hours to complete.

Conclusion

This specific algorithm may not be easily transmittable to real-world applications due to its small sample size. However, the high accuracy, recall, precision, and F1 scores offer a promising possibility that reliable computer algorithms and models could be developed to classify complex literary works such as poetry.

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