Marin Lolic (ejz2sg)

A Quantitative Analysis of William Shakespeare's Sonnets

Introduction

William Shakespeare is primarily known for his plays, but he also composed 154 sonnets. These poems consist of 14 lines[1], usually three stanzas of four lines each, followed by a final stanza of two lines called a couplet. The typical sonnet is approximately 100 words long, and all but one are written in iambic pentameter. Shakespearean scholars have extensively studied the sonnets, and have generally divided them into four categories: "The Fair Youth" (sonnets 1-77 and 87-126), "The Dark Lady" (sonnets 127-152), "The Rival Poet" (sonnets 78-86), and "Anacreontics" (sonnets 153-154). In this paper, I will use the tools of unsupervised machine learning in the natural language processing context to understand the topics and sentiment of Shakespeare's sonnets, and to determine if these support the classic categorization.

Clustering Analysis

I computed a TF-IDF matrix for the terms used in all 154 sonnets, then used K-means clustering with four clusters. The unnormalized TF-IDF data seems to work best, correctly finding many of the patterns in the sonnets. First, it puts sonnets 153 and 154 into their own cluster, recognizing how different they are from the others. It also correctly clusters the majority of the "Rival Poet" sonnets from 78-86, though it also puts some other sonnets in this cluster as well. K-means seems to have the most difficulty differentiating between the "Fair Youth" and "Dark Lady" sequences. While it correctly identifies all the "Dark Lady" sonnets as one cluster, it also adds most of the "Fair Youth" sonnets to the same. This could be due to some thematic overlap between the two groups of sonnets, as the narrator (the "Poet") appears to be in a love

[1] Except Sonnet 99, which has 15 lines

triangle with the Fair Youth and the Dark Lady. Hierarchical clustering (Figure 1) also groups the final two sonnets together.

Principal Component Analysis

Next, I calculated the principal components of the data as well as their loadings from the TF-IDF matrix, looking for words or patterns of words. First, we find that the term that explains the greatest amount of variance in the sonnets is "zealous", while the second through fourth words are some variation of "youth" (Figure 2). The prominence of youth is obvious given that the bulk of the sonnets refer to the "Fair Youth", while the importance of "zealous" is more puzzling. The word only appears once in the corpus (in Sonnet 27) which is about the Poet dreaming of the "Fair Youth." Looking at the loadings of the top two principal components, we find many of the expected themes (Figure 3). The first principal component, PC0, heavily uses second-person pronouns, whether the more modern "you/your" or the Elizabethan "thy/thou/thee." This fits neatly with the majority of the sonnets being directly addressed to the Poet's muse, the "Fair Youth." The second principal component, PC1, combines "hearts/eyes" in the positive direction with "her/she/mistress" in the negative direction. While the former is common in love poetry, the latter almost certainly refers to the "Dark Lady."

Latent Dirichlet Allocation

I next employed LDA to categorize the underlying topics present in Shakespeare's sonnets. To avoid excessively long computation time, I set the number of topics to three, and the number of sampling iterations to 1000. The results are presented in Figure 4. The first topic seems to center on the second person pronouns, much like the first principal component

("thy/thou/thee"). This reflects the personal nature of the sonnets, especially the dominant "Fair Youth" sequences. The second topic continues this trend, but switches the pronouns to more modern forms ("you/your/their"). Finally, the third topic is a mix of various terms with no obvious interpretation. Since the LDA analysis did not use TF-IDF, it seems to have focused heavily on the most frequently-used words, which are not necessarily the most meaningful. It is also possible that more sampling iterations or a different number of topics could have revealed additional information.

Sentiment Analysis

I also analyzed the sentiment of the corpus using the NRC lexicons. Figure 5 shows the results arranged by groups of sonnets. In the "Fair Youth" sequences, the main emotion is joy, followed by trust and sadness, with polarity at the bottom. The "Rival Poet" sonnets see joy and trust remain in the top positions, but now polarity rises into the top 4. The "Dark Lady" sequence sees another large change in sentiment, with anger, fear and disgust rising into positions 3-5. Finally, Sonnets 153 and 154 have their own distinctive feature, as disgust rises to third (nearly second), higher than in any of the other groups. These sentiments fit well with classic characterization of the sonnets: joy and trust are prominent in the majority of cases (especially in regards to the "Fair Youth"), polarity rises when the "Rival Poet" emerges, and negative emotions follow in the "Dark Lady" sequence.

Conclusions

Each of the main methods of analysis revealed something different about the structure and meaning of Shakespeare's sonnets. Clustering uncovered the uniqueness of Sonnets 153-154,

PCA hinted at the importance of the “Dark Lady” sonnets, LDA noted the personal nature of the majority of the sonnets (as most are addressed to a person), and sentiment analysis found clear differences between the “Fair Youth” sequences and the other sonnets, including the “Rival Poet” group. Overall, my analysis lends support to the scholarly consensus, finding some important differences across the corpus while also helping to group the sonnets by theme and sentiment.

Marin Lolic (ejz2sg)

Appendix

Figure 1: Hierarchical Clustering of Sonnets by Euclidean Distance

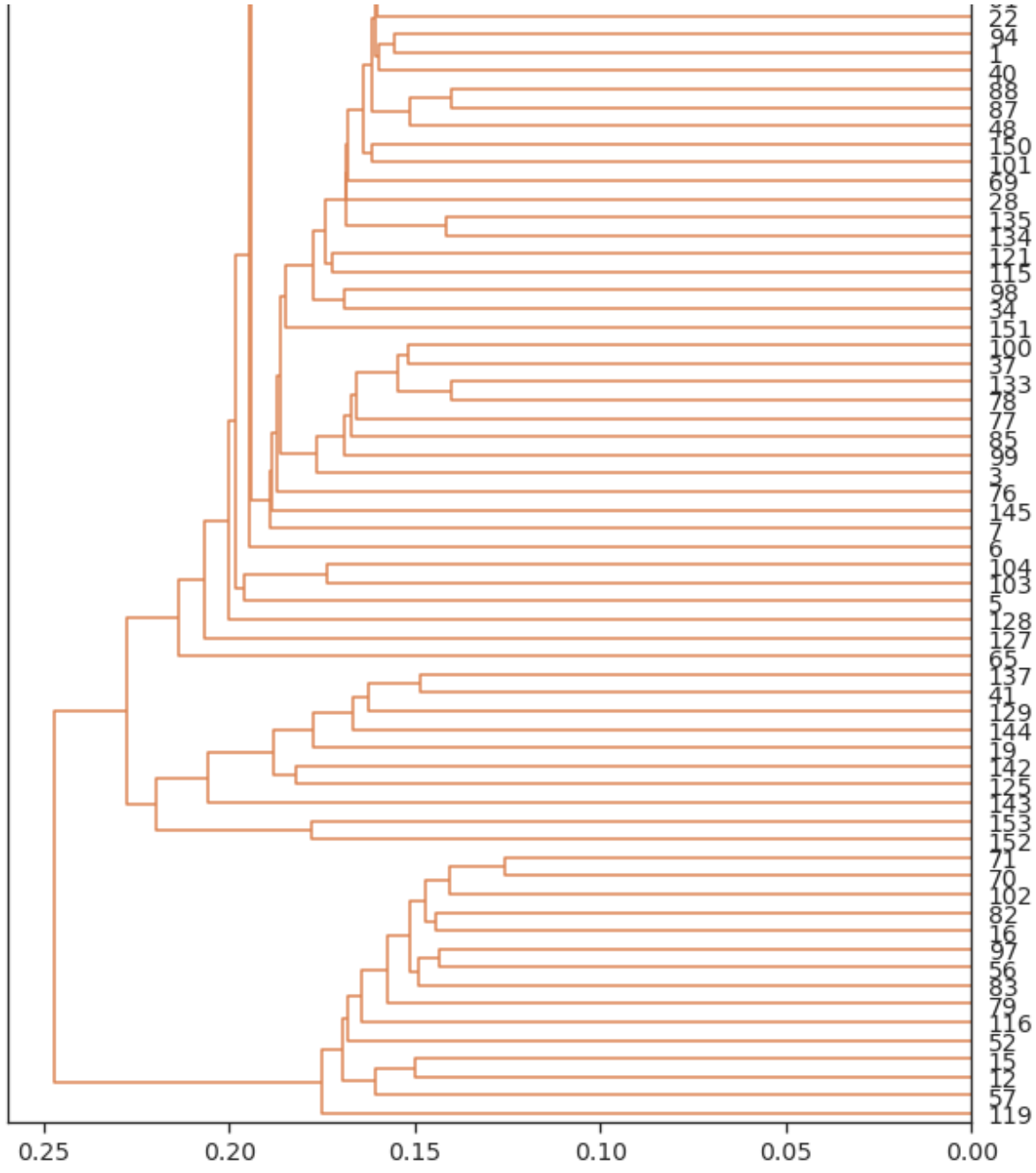
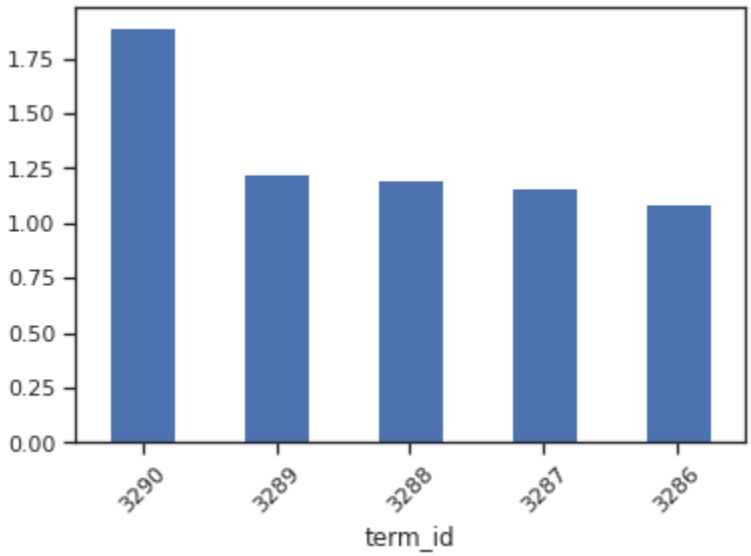


Figure 2: Top 5 Terms by Percent of Variance Explained



term_id	
3290	zealous
3289	youths
3288	youthful
3287	youth
3286	yourselfs

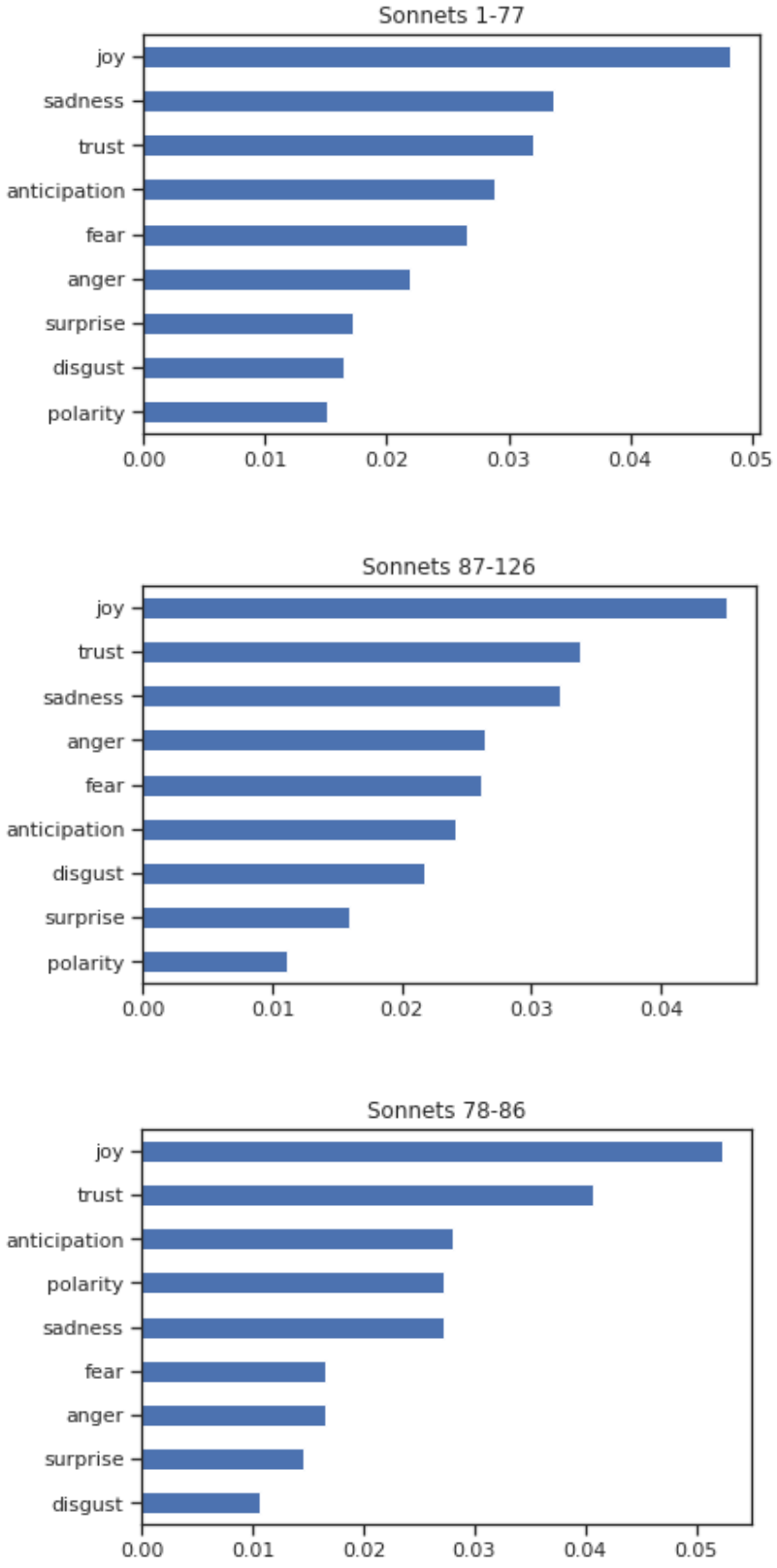
Figure 3: Top Term Loadings by Principal Component

```
PC0+ thy thou thee her heart she thine art dost thyself
PC0- you your yourself should nor world yours were life shall
PC1+ heart eyes mine eye they thy see hearts what doth
PC1- her she mistress loss both sake red fickle will back
```

Figure 4: Topics According to Latent Dirichlet Allocation

```
---------------------------------
topic_id         0 term_str
term_id
2873      0.089030      thy
2852      0.078359     thou
255       0.074692      and
2821      0.054018     thee
1529      0.046349       in
2580      0.024675       so
291       0.017006      art
231       0.016339      all
154       0.015005        a
2840      0.014672    thine
---------------------------------
topic_id         1 term_str
term_id
255       0.058642      and
1529      0.040344       in
154       0.026235        a
3280      0.024691      you
1457      0.021825      his
3283      0.019621     your
1986      0.017416       or
231       0.015653      all
2580      0.015653       so
2823      0.013889    their
---------------------------------
topic_id         2 term_str
term_id
2820      0.041992      the
2889      0.040234       to
1889      0.038281       my
1963      0.036133       of
1499      0.033594        i
2818      0.031445     that
3202      0.017676     with
1192      0.016699      for
1573      0.016406       is
1741      0.016309     love
---------------------------------
```

Figure 5: Sentiment Analysis by Sonnet Group



Marin Lolic (ejz2sg)