Lauren Gulland

SoftDes Section 1

Mini Project 3: Text Mining and Analysis

Project Overview

I used various Shakespeare texts from Project Gutenberg and did simple histogram analysis on them to see which words Shakespeare used more in various works. I hoped to create a grouped bar graph showing the various results.

Implementation My main methods in this project include downloading the plays, pickling them to text files, determining which words are most frequently used in each play, creating a list of the top words from all of the plays, and making a dataframe that presents these findings succinctly. There were multiple ways in which I could have presented this data, but I chose a dataframe not only so I could move forward with attempting to make a bar chart, but also to present the data in a table format that is easy to work with for both the person that is coding the dataframe and the person that is looking at the data.

I have methods that were intended to create grouped bar charts from this data frame as well, but unfortunately could not get these working in the time frame allotted for this project.

Results My results unfortunately do not include a final bar graph like I had hoped, but do instead include a table with all of my results, which you can see below.

```
Top 15 most common words in each play by percentage-frequencies
           Hamlet    Romeo     Lear      Caesar
a        0.312487  0.312288  0.240422  0.210035
and      0.571910  0.463068  0.443672  0.522683
brutus   0.000000  0.000000  0.000000  0.289400
caesar   0.000000  0.000000  0.000000  0.184382
cassius  0.000000  0.000000  0.000000  0.180374
ham      0.211076  0.000000  0.000000  0.000000
i        0.333123  0.348642  0.389712  0.416062
in       0.254117  0.224084  0.178668  0.182779
is       0.199874  0.219317  0.136099  0.208432
it       0.241146  0.000000  0.000000  0.155522
me       0.000000  0.000000  0.136099  0.000000
my       0.302464  0.214549  0.274597  0.000000
not      0.182775  0.167467  0.170874  0.210035
```

```
of       0.393852  0.303348  0.293183  0.306235
that     0.224637  0.215741  0.208646  0.232482
the      0.673911  0.504786  0.512021  0.488212
this     0.175111  0.160912  0.145692  0.000000
thou     0.000000  0.165083  0.000000  0.000000
to       0.430407  0.359965  0.330356  0.334293
with     0.000000  0.177599  0.000000  0.000000
you      0.317793  0.213953  0.272199  0.313449
your     0.000000  0.000000  0.136699  0.000000
```

Some of the interesting results that I pull from this are the differences in usage of "thou" versus "you" and "your", which indicate a large difference in speech – Romeo and Juliet uses thou considerably more than any of the other plays, where the word doesn't even make the top 15 list.

When I expanded the top 15 lists to instead compute for the top 50 words in each play, you get the following result:

```
Top 50 most common words in each play by percentage-frequencies
           Hamlet    Romeo     Lear    Caesar
a        0.312487  0.312288  0.240422  0.210035
all      0.067804  0.069132  0.062354  0.092191
and      0.571910  0.463068  0.443672  0.522683
antony   0.000000  0.000000  0.000000  0.101811
are      0.076648  0.000000  0.081540  0.095398
as       0.129122  0.098335  0.076144  0.112233
be       0.127353  0.139457  0.103124  0.125059
brutus   0.000000  0.000000  0.000000  0.289400
but      0.152116  0.112042  0.079141  0.117043
by       0.066035  0.075688  0.063553  0.092191
caesar   0.000000  0.000000  0.000000  0.184382
casca    0.000000  0.000000  0.000000  0.056918
cassius  0.000000  0.000000  0.000000  0.180374
citizen  0.000000  0.000000  0.000000  0.055315
come     0.058370  0.059001  0.000000  0.059323
did      0.000000  0.000000  0.000000  0.060125
do       0.087261  0.063173  0.065352  0.105819
edg      0.000000  0.000000  0.058757  0.000000
fool     0.000000  0.000000  0.070748  0.000000
for      0.146220  0.148396  0.100126  0.149911
friar    0.000000  0.053637  0.000000  0.000000
from     0.000000  0.060193  0.056358  0.000000
glou     0.000000  0.000000  0.070748  0.000000
```

```
good    0.058960  0.000000  0.000000  0.056918
ham     0.211076  0.000000  0.000000  0.000000
have    0.106128  0.078072  0.124108  0.118646
he      0.125584  0.071516  0.104323  0.154721
her     0.000000  0.092971  0.083938  0.000000
him     0.114972  0.059597  0.118712  0.132274
his     0.175111  0.082840  0.125307  0.128266
hor     0.065445  0.000000  0.000000  0.000000
i       0.333123  0.348642  0.389712  0.416062
if      0.066035  0.061385  0.065951  0.067340
in      0.254117  0.224084  0.178668  0.182779
is      0.199874  0.219317  0.136099  0.208432
it      0.241146  0.143033  0.114515  0.155522
jul     0.000000  0.069728  0.000000  0.000000
kent    0.000000  0.000000  0.103124  0.000000
king    0.113792  0.000000  0.000000  0.000000
know    0.000000  0.000000  0.000000  0.054513
lear    0.000000  0.000000  0.134301  0.000000
lord    0.130891  0.000000  0.058157  0.000000
love    0.000000  0.081648  0.000000  0.000000
me      0.136197  0.157932  0.136099  0.151514
my      0.302464  0.214549  0.274597  0.151514
no      0.083723  0.065557  0.092931  0.074554
not     0.182775  0.167467  0.170874  0.210035
now     0.057781  0.000000  0.000000  0.000000
nurse   0.000000  0.088203  0.000000  0.000000
o       0.067804  0.091779  0.061754  0.057720
of      0.393852  0.303348  0.293183  0.306235
on      0.075469  0.052445  0.059956  0.065736
or      0.065445  0.087607  0.000000  0.000000
our     0.070162  0.000000  0.068949  0.069745
queen   0.069573  0.000000  0.000000  0.000000
rom     0.000000  0.097143  0.000000  0.000000
romeo   0.000000  0.079860  0.000000  0.000000
shall   0.067214  0.066749  0.058757  0.100208
she     0.000000  0.068536  0.000000  0.000000
sir     0.000000  0.000000  0.068349  0.000000
             ...       ...       ...       ...
```

Here, not only do more individual names pop up, but you also start to get some gender-balanced terms: "she" appears in Romeo and Juliet's top 50 but none of the other lists, while "sir" only appears in King Lear's top 50 list. In addition, you can tell that Romeo and Juliet has more exclamatory remarks than all the rest: while "O" appears in each top 50 list, its population is significantly higher in Romeo and Juliet. Other things of note are the difference between the usage of "good" and "fool", which appear in Julius Caesar and King Lear respectively, showing more difference in language and sentiment.

Reflection

When I started this project, I attempted to work with Spotify playlist data, but the API implementation was something that I just couldn't get to work for how much time I was putting into it, so I ended up switching to a project much simpler in scope (at least for collecting data) by using Project Gutenberg and downloading various Shakespeare texts. However, when I tried to get good visualizations and data out of this, I quickly ended up falling into a black hole of learning how to make bar graphs using Bokeh, which led to attempting to learn Panda data frames, so I ended up in the massive time-sink that is trying to get brand new packages and things to work with my code, so I spent wayyyy too long on my project for the unsatisfying lack of data that it produced. I put too many hours into this project honestly, and still haven't gotten anything truly tangible out of it to show other than how much I've learned. It's kind of a shame, and a little frustrating, but I have honestly learned a lot and will probably keep working on this after the project is due because I really want to solve this.