Shakespeare Project

Michael Neuder *

April 2019

1 Introduction

Zipf's Law is a mysterious phenomena present in many literary fields, languages, and contexts. In its most basic form it examines the relationship between the rank of a word in a frequency list, denoted r, and the frequency (number of occurrences in a set of text) of that word, denoted f. The common theme is that the frequency of a word is roughly inversely proportional to the rank [1], or

$$f \propto \frac{1}{r}.$$

Proportional is defined as equivalent up to a constant factor (e.g. $x \propto 5x$ but $x \not\propto x^2$). For example if the most common word appears n times in a large corpus of data, then the second most common word should appear about $n/2$ times and the third most common word should appear about $n/3$ times and so on. The proportionality means that this relationship holds up to some constant factor (e.g. $f = 1/2r$ is still valid because 2 is just a constant factor). This has been shown in many contexts, but why it is true is still unknown. A more general form of this relationship can include a free parameter α which scales the rank,

$$f \propto \frac{1}{r^\alpha}.$$

This gives us more flexibility in our model. One way of verifying this relationship is by plotting frequency and rank on a loglog plot (a normal scatter plot, but with the x and y axis replaced with $\log x$ and $\log y$) [1]. The point of this can be seen by taking the log of both sides of the above proportionality (think of proportionality as super similar to equality, so we can safely take the logarithm of both sides without worry),

$$f \propto \frac{1}{r^\alpha} \implies \log(f) \propto \log(\frac{1}{r^\alpha}).$$

Using the properties of logarithms (namely $\log(x/y) = \log(x) - \log(y)$ and $\log(x^\alpha) = \alpha \log(x)$), we can rewrite this as,

*All code for project can be found here.

$$\begin{aligned}\log(f) &\propto \log(\frac{1}{r^\alpha}) \\ &\propto \log(1) - \log(r^\alpha) \\ &\propto -\alpha \log(r). \quad \text{because } \log 1 = 0\end{aligned}$$

So now we can look at this loglog plot and examine the slope (which is -α). The slope should ideally be somewhere close to -1. Historically this has been done by simply eyeballing the slope of the line on the loglog plot, but instead we will use a simple linear regression to fit the best line to the points and examine this slope (this will be explained clearly in the next section). The goal of this project is to prove that Shakespeare's works deviate from Zipf's Law by showing that the slope of the linear regression of the word frequency plotted against rank is significantly different than the same variable but calculated over modern texts. This will demonstrate that Shakespeare's vocabulary was statistically different (more varied) from some text that we interact with today. To do this we will use a one sample t-test (described in next section). Before we look at some data, we first have to further explore the mathematical theory we will be using.

2 Mathematical Prerequisites

The statistical process we will use to determine if Shakespeare's word usage differs from modern word usage is outlined as follows (this is a lot of technical language but don't worry each step will be explained in detail below),

1. Calculate the slope of the regression line of the loglog plot of modern text and save this value (we will call it μ).
2. For each Shakespeare play, calculate the same slope on the same plot. This will give us a list of slopes which we will call $\mathbf{S} = [s_1, s_2, ..., s_{36}]$
3. Run a one sample statistical t-test to check if $\mathbf{S}$ could be distributed normally with a mean of μ.

2.1 Slope of the Regression Line

Linear regression is the process of fitting a line to a set of data points. The goal is to choose a line such that the average distance of each point from that line is minimized. Let's look at an example. Say we are examining the relationship between high school and college GPA. Below is a scatter plot of what this could look like. In this plot each dot represents a person and their corresponding GPA's.

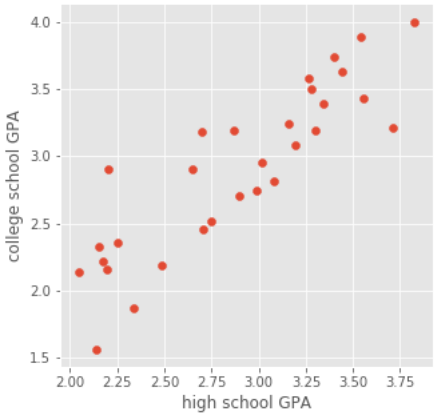

Figure 1: Scatter plot of several individuals.

Clearly there is some correlation, or in other words, high school GPA can predict college GPA to some degree of accuracy. Now linear regression can be used to find a line that fits this data in the "best" way (so that the distance from each point to the line is minimized). Below is a plot with the line of best fit drawn in.

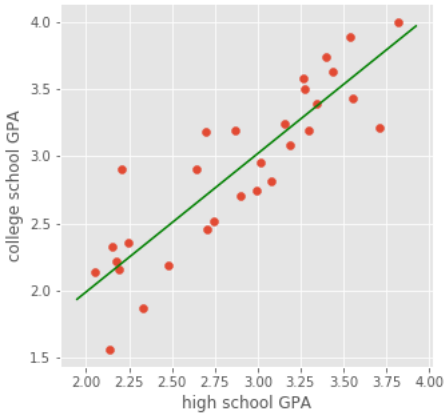

Figure 2: Scatter plot of several individuals with line of best fit.

Now we can examine the slope of this line (in this case it is 1.027), and this is the value of interest. We will do this process 37 times (1 for all the modern text and then 1 for each of the 36 plays). Now that we have all the variables we are interested in, we can conduct the statistical test.

2.2 One Sample t-test

Now we have 37 values (the slopes of all our lines of regression, 1 from the modern text μ and 36 from each of the plays $\mathbf{S}$). Now we want to see if these 36 values are statistically different from the one of the modern text. To do this we will rely on one of the most fundamental objects in statistics; the normal distribution. Almost everything is distributed normally. Below is the canonical plot showing this distribution.

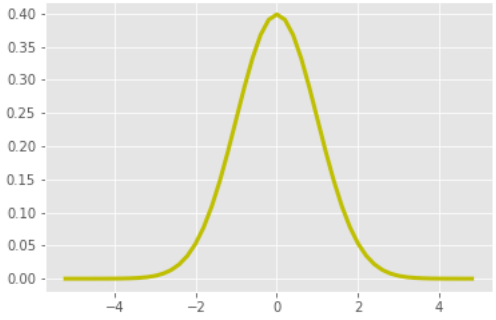

Figure 3: The normal distribution.

We can ignore most of the details of this function and just focus on the intuition. Let's reconsider the information we have. We have our value μ and our set of 36 values $\mathbf{S}$. We want to see if the the mean of our normal distribution (i.e. the highest point – 0 in Figure 3) is μ, then how likely is it that our 36 other variables come from the same distribution (come from is intentionally vague, exactly what this means isn't super important we just want to see if the are statistically similar to our value μ). If this is very unlikely, then we can say that the underlying system that generated our sample of 36 values must be different fundamentally. What this means in the context of Shakespeare is that the variety of language he used in his works is statistically different then the language we interact with today. The whole point of this exercise is to show that Shakespeare's language is distributed differently and thus fundamentally different than today. Thanks for bearing with me through all the theory, now let's look at some of the data.

3 Modern Text

In order to prove Shakespeare's text was more varied than modern text we need a big corpus of current language to examine. We will examine the Wikipedia Movie Plots data set which is available on Kaggle [3]. This data set contains 34,886 movies and contains lots of relevant information about them. What we

want to examine is the summaries. These several sentence descriptions of the stories provide a ton of useful textual data that is similar to stories we interact with daily. Once the data is cleaned we can simply count the number of times each word appears. Below is a table showing the first five most frequent words of this data set.

Rank	Word	Frequency
1	the	752637
2	to	492751
3	and	457061
4	of	357190
5	is	237213

This looks promising, but it is much easier to examine the plots of the data. First we can look at the raw data where the x-axis is the rank and the y-axis is the frequency.

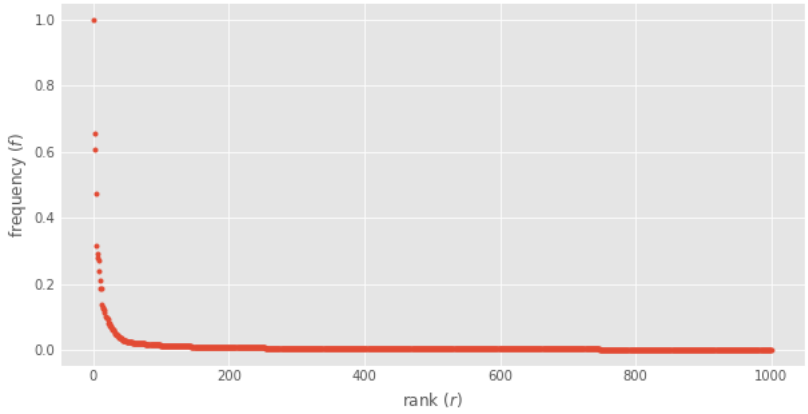

Figure 4: Word frequency as a function of rank for the movie plot summaries.

Now we can also examine the loglog plot of the frequency against the rank.

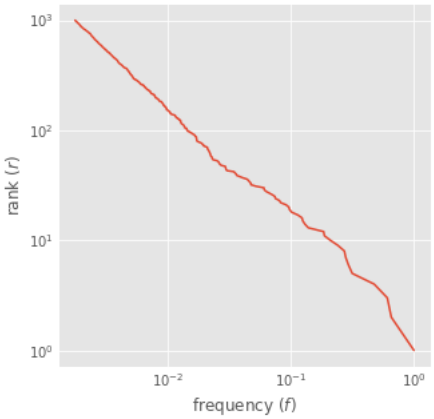

Figure 5: Loglog of frequency against rank.

This clearly gives a negatively correlated relationship with slope around -1 and the exact slope is -0.963 (our μ value). This shows us that the movie summaries are actually less diverse in word usage than Zipf's Law would suggest. This slope is the value we will use to judge how diverse the language is in Shakespeare's plays (e.g. we will use a one sample t-test to see if the α values of the Shakespeare plays are significantly different from this mean). So let's look at some Shakespeare data.

4 Shakespeare Text

The data set we will examine has 36 of Shakespeare's plays (listed in Appendix A) [2]. Each play is stripped of punctuation and the word frequencies are counted. The same process of finding the α value of the loglog plot is performed for each play yielding 36 different plots and slopes. Below is a figure showing a sample of some of these plots.

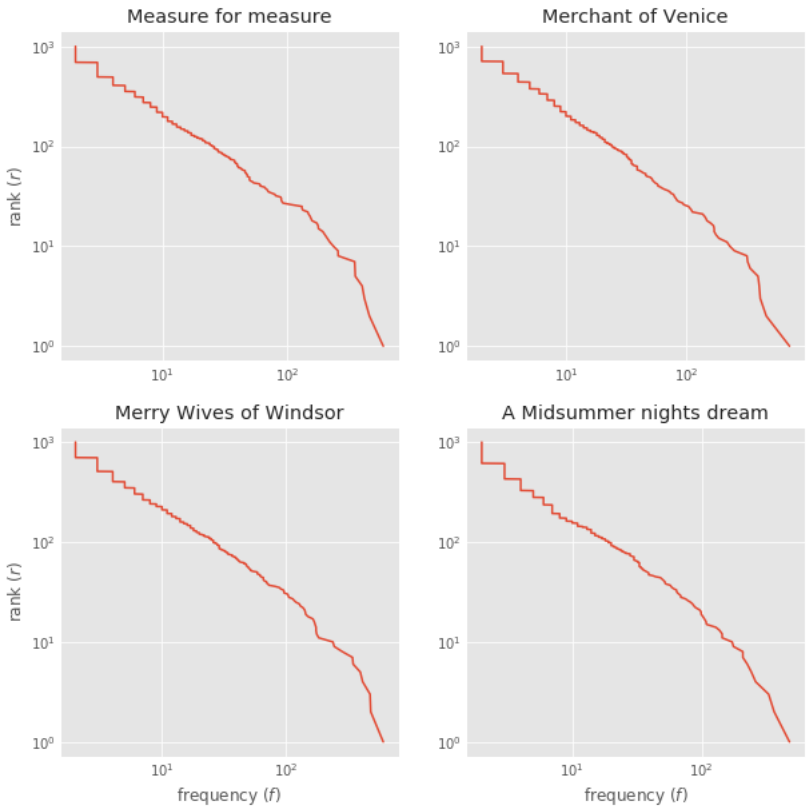

Figure 6: Loglog of frequency against rank for a sample set of Shakespeare plays.

Now each of the α values is collected (full list in Appendix B). We use a one sample t-test to check whether these slopes could belong to a normal distribution with mean -0.963 (from our movie plot data set). This test gives a test statistic of -21.27 which corresponds to a p-value of $1.34x10^{-21}$. Thus we can say with extreme confidence that the the Shakespeare plays differ in word usage in a statistically significant way from the normal text of today.

5 Conclusion

This was all an exercise to prove that Shakespeare used much more varied language than we interact with today. This is relevant because it shows his genius and how complex the literature he wrote was and is another contributing factor

to his legacy. It isn't shocking that his language is much more varied than that of today, but it is cool to show it in a mathematically rigorous way. One other interesting application of this type of analysis could be debates about authorship. Imagine if a new play was found and some people claimed it was written by Shakespeare. Using some more sophisticated statistical methods (but not that different!), we could collect a series of data points for Shakespearean plays and contrast them against other plays of the time period based on this exact process of using the slopes of the regression lines. We could then check if this play in question was written by Shakespeare with high probability. This type of analysis wouldn't be the end all of the debate, but it would give a quantitative perspective to these type of authorship debates. Hopefully this illustrates the interesting perspective quantitative methods can provide to literary field, despite being seen as so distant, and also helps us understand and appreciate Shakespeare's genius just a bit more.

References

[1] Álvaro Corral1 Isabel Moreno-Sánchez1, Francesc Font-Clos. 2016. Large-Scale Analysis of Zipf's Law in English Texts. *PLoS ONE* (2016). https://doi.org/DOI:10.1371/journal.pone.0147073

[2] Liam Larsen. 2017. Shakespeare Plays. (May 2017). Retrieved April 2, 2019 from https://www.kaggle.com/kingburrito666/shakespeare-plays

[3] Justin R. 2018. Wikipedia Movie Descriptions. (March 2018). Retrieved April 2, 2019 from https://www.kaggle.com/jrobischon/wikipedia-movie-plots

Appendices

Appendix A: List of plays used in statistical analysis.

Index	Play	Index	Play
1	Henry IV	19	Measure for measure
2	Henry VI Part 1	20	Merchant of Venice
3	Henry VI Part 2	21	Merry Wives of Windsor
4	Henry VI Part 3	22	A Midsummer nights dream
5	Alls well that ends well	23	Much Ado about nothing
6	As you like it	24	Othello
7	Antony and Cleopatra	25	Pericles
8	A Comedy of Errors	26	Richard II
9	Coriolanus	27	Richard III
10	Cymbeline	28	Romeo and Juliet
11	Hamlet	29	Taming of the Shrew
12	Henry V	30	The Tempest
13	Henry VIII	31	Timon of Athens
14	King John	32	Titus Andronicus
15	Julius Caesar	33	Troilus and Cressida
16	King Lear	34	Twelfth Night
17	Loves Labours Lost	35	Two Gentlemen of Verona
18	Macbeth	36	A Winters Tale

Appendix B: α values for each of the plays

Play	α	Play	α
Henry IV	-1.0798	Measure for measure	-1.1034
Henry VI Part 1	-1.0436	Merchant of Venice	-1.0878
Henry VI Part 2	-1.0526	Merry Wives of Windsor	-1.1184
Henry VI Part 3	-1.0800	A Midsummer nights dream	-1.0131
Alls well that ends well	-1.0914	Much Ado about nothing	-1.1125
As you like it	-1.0972	Othello	-1.1135
Antony and Cleopatra	-1.0824	Pericles	-1.0410
A Comedy of Errors	-1.0849	Richard II	-1.0448
Coriolanus	-1.0842	Richard III	-1.0491
Cymbeline	-1.0531	Romeo and Juliet	-1.0808
Hamlet	-1.0821	Taming of the Shrew	-1.0858
Henry V	-1.0524	The Tempest	-0.9972
Henry VIII	-1.0655	Timon of Athens	-1.0317
King John	-1.0506	Titus Andronicus	-1.0663
Julius Caesar	-1.0788	Troilus and Cressida	-1.0963
King Lear	-1.0809	Twelfth Night	-1.0795
Loves Labours Lost	-1.0495	Two Gentlemen of Verona	-1.0422
Macbeth	-1.0040	A Winters Tale	-1.0989