Onur Talu
2/23/2017

TEXT MINING WRITE UP

Project Overview

I used a word frequency analysis to find the most frequent words in *The Trial* and *The Metamorphosis* by Franz Kafka and tried to relate their frequency relationships to Zipf's Law. I hoped to get a basic understanding of text analysis and getting the information I wanted out of a huge file of data. In the meantime, I learned to use new packages within python that allowed me to draw graphs and represent my work visually.

Implementation

After I used a word frequency analysis, I obtained a list of tuples that contain the word and the number of times it was used in the text. For *The Trial,* the results looks as such:

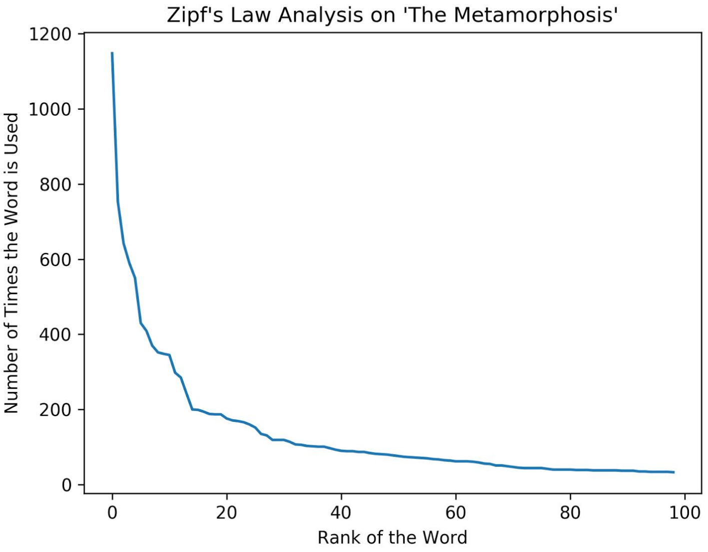
```
[('the', 4727), ('to', 2856), ('he', 2074), ('and', 2006), ('of', 1615), ('it', 1467), ('that', 1431), ('a',
1320), ('you', 1268), ('in', 1244), ('i', 1223), ('was', 1190), ('k', 1176), ('his', 1014), ('s', 975), ('as',
912), ('had', 811), ('said', 770), ('him', 750), ('but', 737), ('not', 679), ('with', 665), ('at', 614), ('for',
611), ('be', 606), ('on', 534), ('this', 527), ('t', 525), ('have', 511), ('there', 495), ('they', 464), ('if',
446), ('would', 442), ('is', 425), ('so', 424), ('been', 387), ('all', 384), ('me', 370), ('what', 366),
('from', 365), ('no', 342), ('about', 341), ('them', 336), ('then', 332), ('now', 312), ('out', 312),
('even', 308), ('could', 306), ('do', 305), ('she', 298), ('her', 296), ('who', 284), ('were', 281), ('by',
277), ('up', 268), ('one', 263), ('can', 263), ('more', 257), ('very', 248), ('just', 245), ('time', 245),
('when', 244), ('only', 244), ('are', 227), ('some', 218), ('lawyer', 214), ('like', 212), ('did', 210),
('here', 209), ('way', 208), ('asked', 199), ('door', 198), ('my', 197), ('man', 195), ('room', 193),
('which', 187), ('himself', 187), ('re', 185), ('ve', 184), ('any', 184), ('don', 183), ('see', 181),
('down', 179), ('than', 172), ('m', 170), ('or', 168), ('an', 167), ('how', 165), ('court', 163), ('your',
161), ('go', 161), ('hand', 161), ('over', 157), ('other', 153), ('much', 153), ('still', 151), ('back',
149), ('into', 148), ('looked', 144), ('painter', 141)]
```

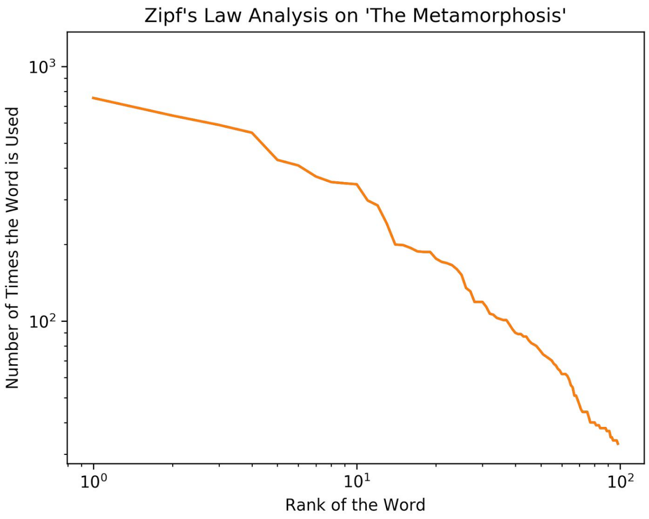
While, for *The Metamorphosis*, the results looks like such:

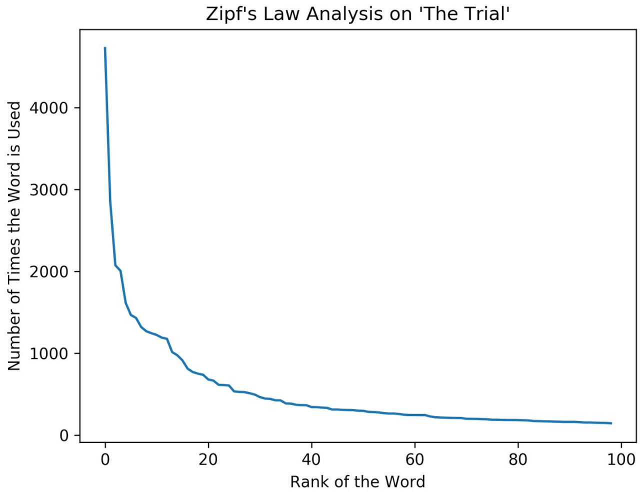
```
[('the', 1148), ('to', 753), ('and', 642), ('he', 590), ('his', 550), ('of', 430), ('was', 409), ('it', 370),
('had', 352), ('in', 348), ('that', 345), ('gregor', 298), ('a', 285), ('as', 242), ('she', 200), ('with', 199),
('s', 194), ('him', 188), ('would', 187), ('her', 187), ('not', 176), ('but', 171), ('at', 169), ('for', 166),
('they', 160), ('on', 152), ('all', 135), ('room', 131), ('from', 119), ('be', 119), ('could', 119), ('out',
114), ('have', 107), ('there', 106), ('if', 103), ('father', 102), ('been', 101), ('sister', 101), ('so', 97),
('this', 93), ('i', 90), ('mother', 89), ('now', 89), ('door', 87), ('himself', 87), ('then', 84), ('back', 82),
('up', 81), ('even', 80), ('into', 78), ('what', 76), ('no', 74), ('did', 73), ('one', 72), ('more', 71), ('their',
70), ('when', 68), ('were', 67), ('about', 65), ('them', 64), ('t', 62), ('you', 62), ('way', 62), ('only',
61), ('time', 59), ('by', 56), ('than', 55), ('said', 51), ('just', 51), ('little', 49), ('any', 47), ('other', 45),
('still', 44), ('do', 44), ('first', 44), ('get', 44), ('or', 42), ('go', 40), ('while', 40), ('made', 40), ('some',
40), ('without', 39), ('see', 39), ('again', 39), ('after', 38), ('much', 38), ('like', 38), ('before', 38),
```

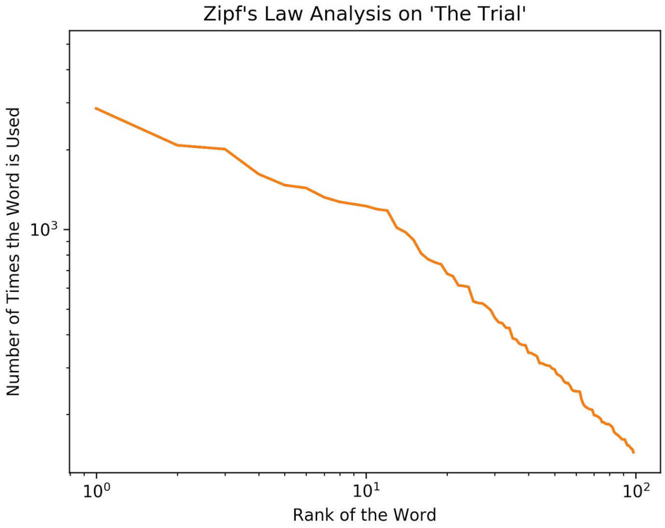
('head', 38), ('where', 37), ('clerk', 37), ('chief', 37), ('down', 35), ('open', 35), ('we', 34), ('very', 34), ('samsa', 34), ('which', 34), ('who', 33), ('over', 32)]

Even though looking at the most frequent words (other than the expected ones, such as ‘the’, ‘to’ or ‘and’) reveal a lot about the texts (‘he’ is really abundant in both texts, since both books have a strong focus on the conflicts of the protagonist and is told from an omniscient point of view), what I found to be more interesting was to relate the outcome to Zipf’s law. Zipf’s law states that the of frequencies of each word in a long text, will be with proportion to the inverse of the rank of the word. Generally, if the logarithm of the plot is taken, something similar to a straight line should be obtained.









As seen in the figures above, when the logarithms of the plots are taken and when the text length is increased (*The Trial* is almost three times as long as *The Metamorphosis*), the trend looks closer to the Zipf's law.

Reflection

I worked a lot on this project and it took me a lot to get word frequency analysis working. Other than that, it took me a while to stip the list off of the header comments and the license; and getting the matplotlib to work. I took a lot of help of NINJAs and friends, and if it weren't for them, I wouldn't be able to finish the project. Still, I would really like to iterate on this project, create word clouds and do analysis on more texts than just two, for Mini-Project 5.