201505547
SHIVAM KHARE

First I tokenize the text of 5 different authors form different time-span and for each author I calculate the bigrams and unigrams and plot the log of frequency vs log of rank. But to compare different texts of varying length and frequency, we need to normalize the texts by taking equal size corpus.

Observation - Graph was nearly straight line. For each author It was clearly visible that every author has its own set of dictionary and that defines their unique style of writing.

For smoothing , I applied plus 1 to numerator and divided by vocab size. This method is called Laplace smooting. For testing this, I am asking user to enter any sentence and finding probab of entered sentence.
Graph was generated and is in folder .

Graph observation -

As can be seen, the match is good but with divergences at high and low ranks. The most frequent words at the top left of the graph are below the expected line which means they are less common than they should be. The less frequent words are also less common than they should be. Sometimes they are more common than they should be.

In addition to local slope variations at high and low ranks, the overall slope is also observed to vary around 1 again with variations hypothesised to be caused by author differences.

If many texts are examined, all obey the law to a large degree but there are differences that are ascribed to age,educational attainment, audience, gender, linguistic ability where the written language is not the mother tongue of the writer and general mental health .

Given text, can we predict the author?

Zipf distance

One approach to compare two texts is to look at how the same words are ranked between the texts. By arranging words in order of frequency so that the most frequent is ranked first,two texts can be lined up against one another and the differences between ranks assessed to determine a distance. If it is assumed that the rank order of words is a characteristic of a particular author, any differences would indicate that two texts have different authors.

This approach is taken by Havlin (1995) where a distance measure is defined that considers common words (common in this context means common between the texts) and calculates the square root of the mean sum of the squared differences in rank between all individual words that appear in both texts. Mathematically, this is expressed in equation 2.3.1 which shows the distance between two texts.

r12 = [(1/N) summation(r1(λ) − r2(λ))^2]^(0.5)

Where r12 is the rank distance between the texts, N is the number of common words individually denoted by λ, r1(λ) and r2(λ) are the ranks of the same word in the two texts.

A maximum number of words is chosen as a parameter. The rank-frequency constant that Zipf’s law predicts is not included in the distance calculation. Initial experiments by Havlin (1995) and subsequently by Vilensky (1996) showed that texts written by the same author are closer to one another than texts written by different authors. The distance measure requires that only words shared by both texts can be considered.This has the effect that less common words, the content words, would tend to be eliminated and function words only would tend to be favoured. Content words in this context means words that give information about the content of the text which tend in general not to be shared between texts. Function words are the more frequent words that act to join content words together and which are more common and therefore more likely to be shared between

texts.
Given the result that works by the same author tend to be closer to one another, this 2.3. AUTHORSHIP ATTRIBUTION 22 provides evidence that individual authors have a characteristic hierarchy of function words. It also shows that divergences from Zipf's law can be exploited for author attribution. The frequency of individual words is not considered in the distance calculation. For relatively low ranked words, the rank could vary considerably with small variations in word frequency and is likely to mean that increasing the number of words used will lead to big variations in the distance measure. This underlines where the measure is likely to prove useful for frequent function words. For these words, it may be that the frequency information would give additional subtle information that could be used to distinguish between texts.

2. Word frequency

3. Rank order

4. Vocabulary richness
In contrast to text structure, Zipf's law, in its frequency spectrum form, is able to make predictions about some features of vocabulary richness (Montemurro and Zanette, 2002) such as the number of unique words, the ratio of unique words to total words as well as the number of words that appear only once.

The Zipf distance technique discussed considers differences in rank
but does not take account of the frequency of words and this may cause it to lose predictive power particularly for less frequent words.

The rank order technique appears to combine rank and frequency information and on
that basis would be expected to outperform the Zipf distanc.