Topics in Data Management

Zipf's Law

Date: 02/05/2017

Author: Chaitali Sudam Kamble (csk3565)

What does Zipfs Law say [1]?

The Zipfs Law is used for text processing and it states that the rank of a word is inversely proportional to its frequency, and the multiplication of rank and probability is a constant value and can be effectively explained as, the most frequent word in file has doubled the frequency than that of second frequent word in a file and so on.

Text Used for Zipfs Law [2]:

The text which is used for Zipfs Law is a file consisting of the information of the great cricketer Sachin Tendulkar and his achievements while serving his country for 20 long years. He is one of the legends of the game called Cricket and known as 'Master Blaster'. His contributions to the cricket and his popularity in this game is massive and appreciable. Apart from a player, he is a very nice and humble human being and well known for his down to earth nature even if he already achieved all success in the game. If cricket is a religion, Sachin is known as the God. His contributions towards this game lead him to win almost all awards from the government of India as well as all over the world.

Functioning of Python code:

The submitted python code ZipfsLaw.py takes a text file as an input.

Data cleaning and preprocessing activities done:

1. Removed white spaces from the text in a file.
2. Removed special characters from the text in a file.
3. Converted all the words to lower case to avoid mismatching of lower and upper cases later on.

Maintained a dictionary to track the counts for unique words from a text file. The output generated from this function is a chart showing top 10 occurring words, frequencies, probabilities and the product of rank probability calculations.

```
Zipfs Law results:
| WORD:  the | Frequency:  39 | Rank:  1 | Probability:  0.08590308370044053 | rPr:  0.08590308370044053
| WORD:  of | Frequency:  20 | Rank:  2 | Probability:  0.04405286343612335 | rPr:  0.0881057268722467
| WORD:  in | Frequency:  19 | Rank:  3 | Probability:  0.04185022026431718 | rPr:  0.12555066079295155
| WORD:  and | Frequency:  14 | Rank:  4 | Probability:  0.030837004405286344 | rPr:  0.12334801762114538
| WORD:  to | Frequency:  11 | Rank:  5 | Probability:  0.024229074889867842 | rPr:  0.1211453744493392
| WORD:  his | Frequency:  9 | Rank:  6 | Probability:  0.019823788546255508 | rPr:  0.11894273127753305
| WORD:  he | Frequency:  9 | Rank:  7 | Probability:  0.019823788546255508 | rPr:  0.13876651982378857
| WORD:  tendulkar | Frequency:  6 | Rank:  8 | Probability:  0.013215859030837005 | rPr:  0.10572687224669604
| WORD:  was | Frequency:  5 | Rank:  9 | Probability:  0.011013215859030838 | rPr:  0.09911894273127754
| WORD:  first | Frequency:  5 | Rank:  10 | Probability:  0.011013215859030838 | rPr:  0.11013215859030838
```

Figure 1.0: Results after running ZipfsLaw.py python file.

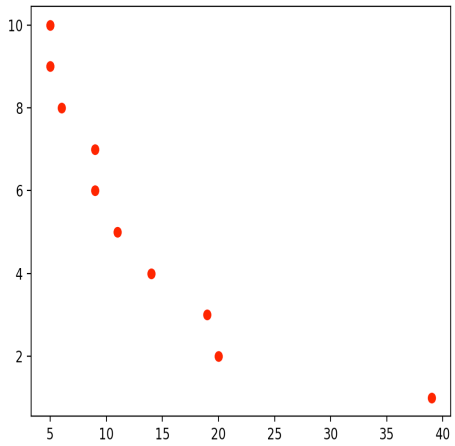

Figure 2.0: A graph generated from python file where x-axis represents frequencies and y-axis represents ranks of words from the file

How the text used conforms the Zipfs Law?:

According to the generated chart from ZipfsLaw.py python file, following are the observations:

1. Zipf's Law fits perfectly on the large text but not on small text.
2. For small text, the frequencies of some words are almost equivalent. For example, the top most occurring word does not have the frequency which is thrice that of the third occurring word.
3. The multiplication of rank and probability of each word from the top occurring words is approximately a constant. Thus, we can say that rank of any word is inversely proportional to its frequency (from the figure 2.0).
4. From the text considered, Zipf's Law does not 100% fit for the provided data.

References:

1. "Zipf's Law, From Wikipedia, the free encyclopedia". Retrieved from: https://en.wikipedia.org/wiki/Zipf's_law. Accessed on: 02/05/2017.
2. "Sachin Tendulkar, From Wikipedia, the free encyclopedia". Retrieved from: https://en.wikipedia.org/wiki/Sachin_Tendulkar. Accessed on: 02/05/2017.