A-9. TEXT ANALYSIS BY MEANS OF ZIPF LAW

Alina Andersen, Giedrė Būdienė, Alytis Gruodis

Vilnius Business College, Kalvariju street 125, Vilnius, Lithuania
email: alina@kolegija.lt

Keywords: Zipf law; Shenon function; ZS-dependence.

Formalized human language analysis and recognition is considered to be a very difficult task in programming history. Formulated in pioneering time of computing (in the middle of XX century) conversation between human and robot has remained as the unsolved problem which has shifted into artificial intelligence concept with the task still to be solved.

Stochastic / chaotic nature of human speech requires the basics of statistical analysis. Zipf law [1] uncovers the relationship between word frequency $\omega(r)$ and its rank r when finite number of set N (size of corpus) is prelimited:

$$Z_1 = \frac{r * \omega(r)}{N} \quad (1)$$

So called Zipd constant Z_1 approximately is stable for infinite corpus belonging to defined language.

By analyzing the frequency of items (words, bywords, ..., n-grams, etc) in the logically related sets (novels, encyclopedias, books, etc) Zipf constant could be treated as a parameter of corpus recognition.

We decided to analyze the item distribution as a distributions grounded on chaotic behaviour. In that case, Shenon function represents an criterion of order / chaos which may be useful for analysis:

$$H(X) = -\sum_{r=1}^{N} p(x_r) * log_2[p(x_r)] \quad (2)$$

We have selected the classical novel by Oscal Wilde [2] as an object of investigation. Fig. 1 represents Log-Log distribution - word frequency dependence on a rank of items. This distribution is typical for small size corpus (in our case it is about 60000 items), when long tail is placed out of direct line [3]. Fig. 2 represents Shenon distribution on a rank of items. Combining such two functions (Shenon function as argument x, Zipf function as argument y) we decided to look into this complicated dependence, so called ZS-dependence (see Fig. 3.) The integrated behaviour of slope is obvious, excluding several chaotic places of grand importance. Such two places corresponds to the set of items when morphologically different content is present.

Presented novel method based on analysis of ZS-dependence requires wide and thorough testing of different corpora.

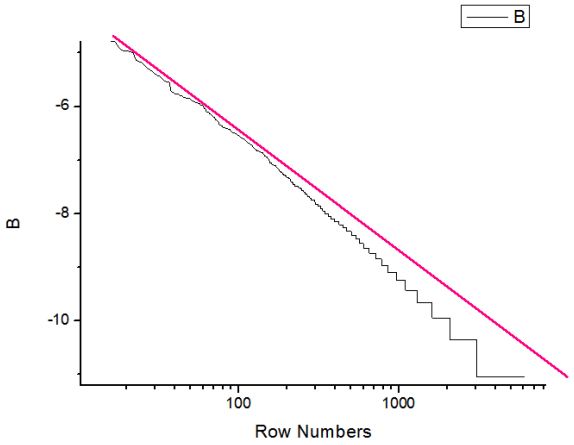

Fig. 1. Log-Log distribution - word frequency dependence on a rank of items.

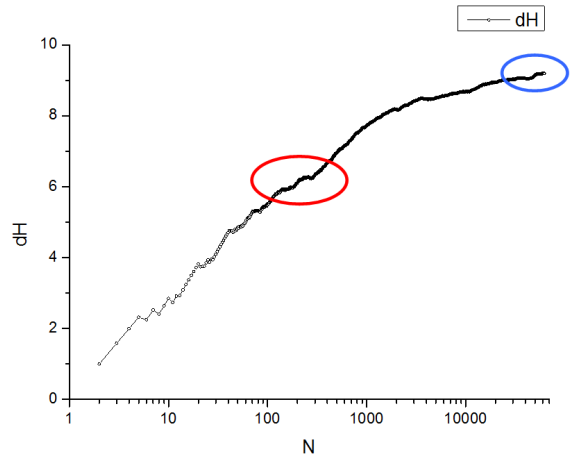

Fig. 2. Shenon distribution on a rank of items.

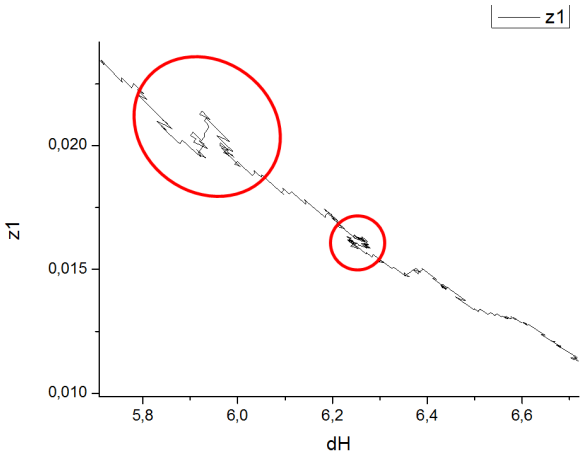

Fig. 3. ZS-dependence.

References

1. G. K. Zipf. Human Behavior and the Principle of Least Effort – Addisson-Wesley, Cambridge, MA, 1949.
2. Oscar Wilde. The Picture of Dorian Gray – Free access Gutenberg library <www.gutenberg.org>.
3. L. Egghe. The dependence of the height of a Lorenz curve of a Zipf function on the size of the system – *Mathematical and Computer Modelling* 43 (2006) 870-879.