Report of Deep Learning for Natural Langauge Processing

Shenghua Han
Sy2303405
419130449@qq.com

Abstract

This is a Report of Deep Learning for Natural Langauge Processing. The natural language is a kind of information who is too difficuly to procress with simple method. So it is necessary to abstract some factors describe the sentence. As a kind of language, Chinese also needs such factors. So the Zipf's law and the entropy will be mentioned in this report.

Introduction

To research a kind language, it is necessary to get the information hidden in the sentence. There are more than ten thousand characters in Chinese. So several thousands words will be used in the daily life.

Zipf's Law mainly talks about the relationship betweem one word's frequency in a sentence and its frequecy ranking in all words mentioned in the sentence. Actually, when a word is seldom mentioned, it will include more information. On the other hand, some word used most frequently may have no concrete meaning. So they are called stopwords. With stopwords, the concrete words will joined up to make up a sentence. What'more, it is possible to get the useful words in the sentence by get the mainly stopwords.

Entropy of information will show the uncertainty of information. So it can be used to think about some characteristics of the information from the angle of the machine. What's more, different words will have some relationship what makes some words will be presented together more easily. So there would be a new factor, joint entropy. We think about the binary joint entropy and the three-way joint entropt.

Methodology

There are models of my research. To verify the Zipf's law, I take the frequency of the useful words and get the graph about the frequency and the sequence in all words. As a result, the Zipf's Law is verified basicly.

As for the entropy of information, it is got by the formula and the frequency of all the word.

The stop word was processed with the 'jieba' in the python.

Experimental Studies

The figures shows it works so well

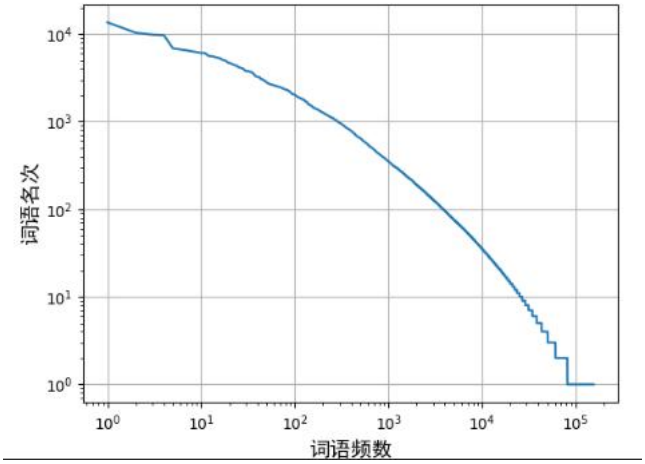

Figure 1： this is the very graph of Zipf'law(based on the words from sixteen novels written by JinYong)

Table 1: the result of entropy of information(based on the characters from '雪山飞狐.txt')

Entropy	9.263333197621813
Number of Characters	408055

Table 2: the result of entropy of information(based on words from sixteen novels written by JinYong)

	Indepental entropy	Binary	Three-way
Entropy	12.01314	6.89149	2.4166
Number of Words	4430791	4430774	4430757

Conclusions

By good scientific writing in right format, it can express you idea of research clearly to others, therefore, we need to learn how to write a good report and even a paper.

References

[1] Zenchang Qin and Lao Wang (2023)， How to learn deep learning? Journal of Paper Writing,

Vol. 3: 23: pp. 1-12.