Report of Calculation of Chinese Information Entropy

Zhenyang Zhu
zhenyangzhu@buaa.edu.cn

Abstract

This is the first assignment report for the NLP course. In this report, the information entropy of Chinese text is calculated, using a database from Jin Yong's novels. This report calculates the entropy of uni-gram, bi-gram and tri-gram language models respectively. At the same time, the report also considers the effect of stop words when calculating information entropy, and makes a comparison.

Introduction

Entropy is generally a measure of the state of some material system, the degree to which certain material system states are likely to occur. The concept of entropy was proposed by the German physicist Clausius in 1865 and has been widely used in thermodynamics. Information entropy is a basic concept of information theory, which is used to describe the uncertainty of the occurrence of each possible event of an information source. In the 1940s, Shannon, the father of information theory, referred to the concept of thermodynamics, called the average amount of information after eliminating redundancy——"information entropy", and gave a mathematical expression to calculate information entropy.

The definition of information entropy is as follows:

$$H(X) = -\sum_{x \in \mathrm{X}} p(x) \log p(x) \tag{1}$$

And it says: $0 \cdot \log(0) = 0$.

There are three properties of information entropy:

1. Monotony: the higher the probability of an event, the lower the information it carries;

2. Non-negative: information entropy can be regarded as a breadth quantity, and non-negative is a reasonable necessity;

3. Summation: The measure of total uncertainty existing in the simultaneous occurrence of multiple random events can be expressed as the sum of measures of uncertainty of each event, which is also an embodiment of breadth.

Shannon strictly proved mathematically that a random variable uncertainty measure function satisfying the above three conditions has a unique form:

$$H(X) = -C\sum_{x\in X} p(x)\log p(x) \tag{2}$$

where C is a constant, the information entropy formula is obtained by normalizing it to C = 1.

Methodology

In this report, the n-gram language model is used to process natural language text. Language model is used to calculate the probability of a sentence, that is, the probability of determining whether a sentence is spoken by a person. Given a sentence:

$$S = W_1, W_2, \ldots, W_n$$

Its probability can be expressed as:

$$P(S) = P(W_1, W_2, \ldots, W_n) = p(W_1)p(W_2 \mid W_1)\ldots p(W_n \mid W_1, W_2, \ldots W_{n-1}) \tag{3}$$

Or written as:

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1}) \tag{4}$$

However, the above formula cannot be directly used to calculate the conditional probability, because it will lead to too large parameter space and seriously sparse data. When the amount of text is not large enough, too much conditional probability will make the final result of each item tend to 0.

Thus, Markov assumption is introduced: the occurrence probability of a random word is only related to a limited number of words before it. Then we get the previous probability calculation simplified as follows:

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1}) \tag{5}$$

Let k = 0 in (5), then the model is called Uni-gram model, that is, w_i is not related to any word, and each word is mutually independent. P(W) is calculated as follows:

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i) \tag{6}$$

Let k = 1 in (5), the model is called Bi-gram model, which means w_i is only related to the first word before it. Then, P(W) is calculated as follows:

$$P(S) = P(w_1 w_2 \ldots w_n) \approx P(w_1)P(w_2 \mid w_1)\ldots P(w_n \mid w_{n-1}) \tag{7}$$

And by the same token, let k = 2, which is called Tri-gram. And when k = n-1, the model becomes an n-element model: N-grams.

In addition, there are many words in Chinese that have no actual meaning, such as "了", "呢", "也" and so on. These words are called stop words.They are extremely common, and as a result, these words rarely express the core information of a document alone. These stop words are seldom helpful if each word rather than a phrase is considered in the retrieval process. Therefore, when calculating the entropy of different language models, this report considers the differences under three conditions: all texts are taken into account, only punctuations are deleted and delete all stop words.

Experimental Studies

Table 1 shows the information entropy of Chinese text in different language models when considering full text, deleting only punctuation and deleting stop words. It can be seen that the information entropy of the text decreases as the value of N increases. This is because the larger the value of N, the simpler the distribution of phrases obtained by word segmentation. The larger N is, the more fixed words there are. Fixed words can reduce the chance of being disrupted by words or short words, make the article more orderly, and reduce the uncertainty of words and sentences formed by words. In other words, the information entropy of the text is reduced, which conforms to the actual cognition.

Table 1 The information entropy of Chinese text in different language models

	Full Text	Punctuaion Deleted	Stop Words Deleted
Uni-gram	10.72	12.16	13.59
Bi-gram	6.57	6.96	6.53
Tri-gram	3.25	2.31	1.18

Table 2 shows the top 10 words that appear frequently under three different text processing modes of different language models. As you can see, punctuation appears most frequently when all texts are considered. When punctuation is removed, the mort frequent words became stop words which are meaningless. When the stop words are removed, the remaining words become the most used words in Jin Yong's novels.

Table 2 The top 10 words that appear most frequently in different language models

	Full Text		Punctuaion Deleted		Stop Words Deleted	
Uni-gram	1.， 2.。 3.的 4.： 5.了	6.“ 7.” 8.他 9.是 10.道	1.的 2.了 3.他 4.是 5.道	6.你 7.我 8.在 9.也 10.这	1.道 2.说 3.便 4.中 5.说道	6.听 7.见 8.韦小宝 9.一个 10.一声
Bi-gram	1.：“ 2.道： 3.。” 4.？” 5.了。	6.，你 7.说道： 8.：「 9.，我 10.!”	1.道你 2.叫道 3.道我 4.笑道 5.听得	6.都是 7.了他 8.他的 9.也是 10.的一声	1.笑道 2.韦小宝道 3.甚麼 4.忽听 5.低声道	6.道说 7.令狐冲道 8.大声道 9.突然间 10.站起身
Tri-gram	1.道：“ 2.说道：“ 3.，说道： 4.：“你 5.道：「	6.：“我 7.了。” 8.叫道： 9.，道： 10.笑道：	1.只听得 2.忽听得 3.站起身来 4.吃了一惊 5.哼了一声	6.笑道你 7.啊的一声 8.点了点头 9.叹了口气 10.说到这里	1.韦小宝笑道 2.五岳剑派 3.砰砰乱跳 4.叹口气说道 5.〖新语丝电子	6.新语丝电子文库 7.叹口气道 8.站起身说道 9.说甚麼 10.----〖新语丝

Conclusions

In this report, the N-Gram model (n=1,2,3) is established for the data set of Jin Yong's novels, and on this basis, the information entropy of Chinese is calculated. At the same time, this report also analyzes the effects of punctuation and stop words on information entropy. The experimental results show that in the N-Gram model, the larger N is, the more different words are in the thesaurus, and the lower the information entropy of the text is. In addition, for uni-gram model, the information entropy increases significantly when punctuation marks and stops are excluded, and decreases when N increases due to fewer words and more dispersed distribution.

References

[1] Brown, P.F., et al., An Estimate of an Upper Bound for the Entropy of English. Computational linguistics - Association for Computational Linguistics, 1992. 18(1): p. 31-40.