## 基于语料库的文本信源的信息论研究与应用实验目的

信源来自：维基百科的资料

## 信源文档

Those who came before us made certain that this country rode the first waves of the i ndustrial revolution, the first waves of modern invention, and the first wave of nuclear pow er, and this generation does not intend to founder in the backwash of the coming age of s pace. We mean to be a part of it--we mean to lead it. For the eyes of the world now loo k into space, to the moon and to the planets beyond, and we have vowed that we shall no t see it governed by a hostile flag of conquest, but by a banner of freedom and peace. We have vowed that we shall not see space filled with weapons of mass destruction, but with instruments of knowledge and understanding.

Yet the vows of this Nation can only be fulfilled if we in this Nation are first, and, th erefore, we intend to be first. In short, our leadership in science and industry, our hopes fo r peace and security, our obligations to ourselves as well as others, all require us to make this effort, to solve these mysteries, to solve them for the good of all men, and to become the world's leading space-faring nation.

We set sail on this new sea because there is new knowledge to be gained, and new ri ghts to be won, and they must be won and used for the progress of all people. For space science, like nuclear science and all technology, has no conscience of its own. Whether it will become a force for good or ill depends on man, and only if the United States occupie s a position of pre-eminence can we help decide whether this new ocean will be a sea of peace or a new terrifying theater of war. I do not say that we should or will go unprotect ed against the hostile misuse of space any more than we go unprotected against the hostile use of land or sea, but I do say that space can be explored and mastered without feeding the fires of war, without repeating the mistakes that man has made in extending his writ around this globe of ours.

There is no strife, no prejudice, no national conflict in outer space as yet. Its hazards are hostile to us all. Its conquest deserves the best of all mankind, and its opportunity for peaceful cooperation many never come again. But why, some say, the moon? Why choose this as our goal? And they may well ask why climb the highest mountain? Why, 35 years ago, fly the Atlantic? Why does Rice play Texas?

目标文本：是从信源单独截取出来的一段test文本：在test\_sample.txt

具体为：

But why, some say, the moon? Why choose this as our goal?

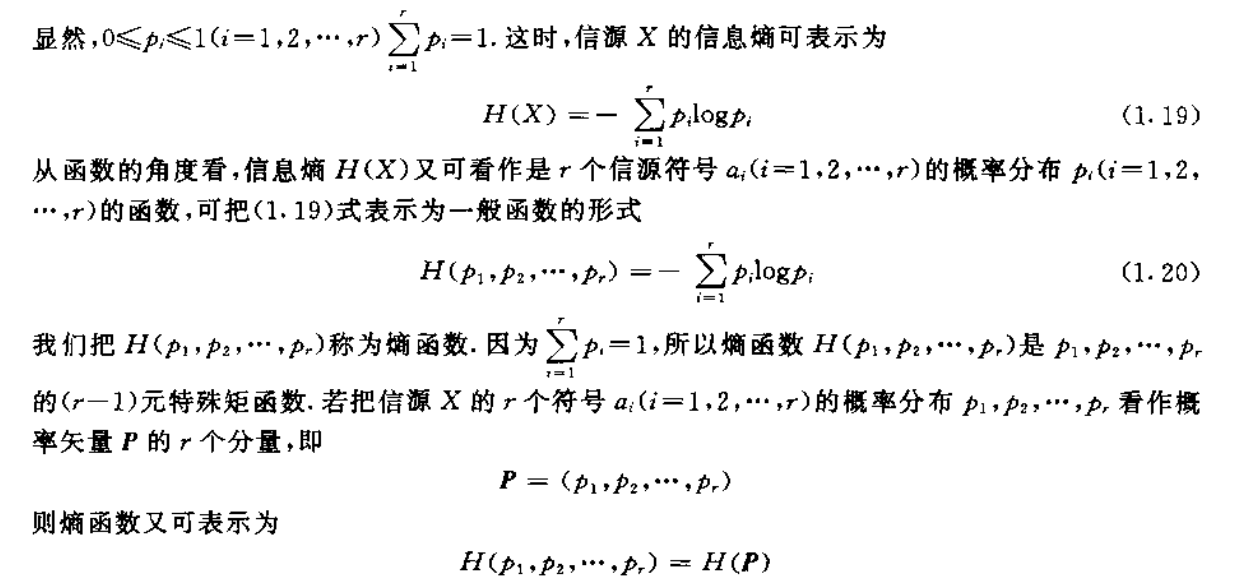
And they may well ask why climb the highest mountain? Why, 35 years ago,

fly the Atlantic?

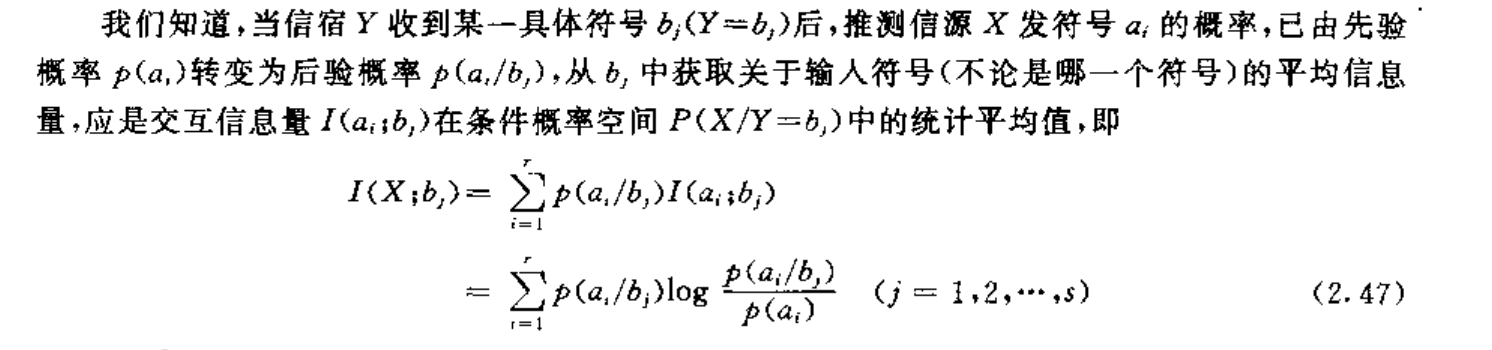
Why does Rice play Texas?

## 理论依据：

熵函数计算：

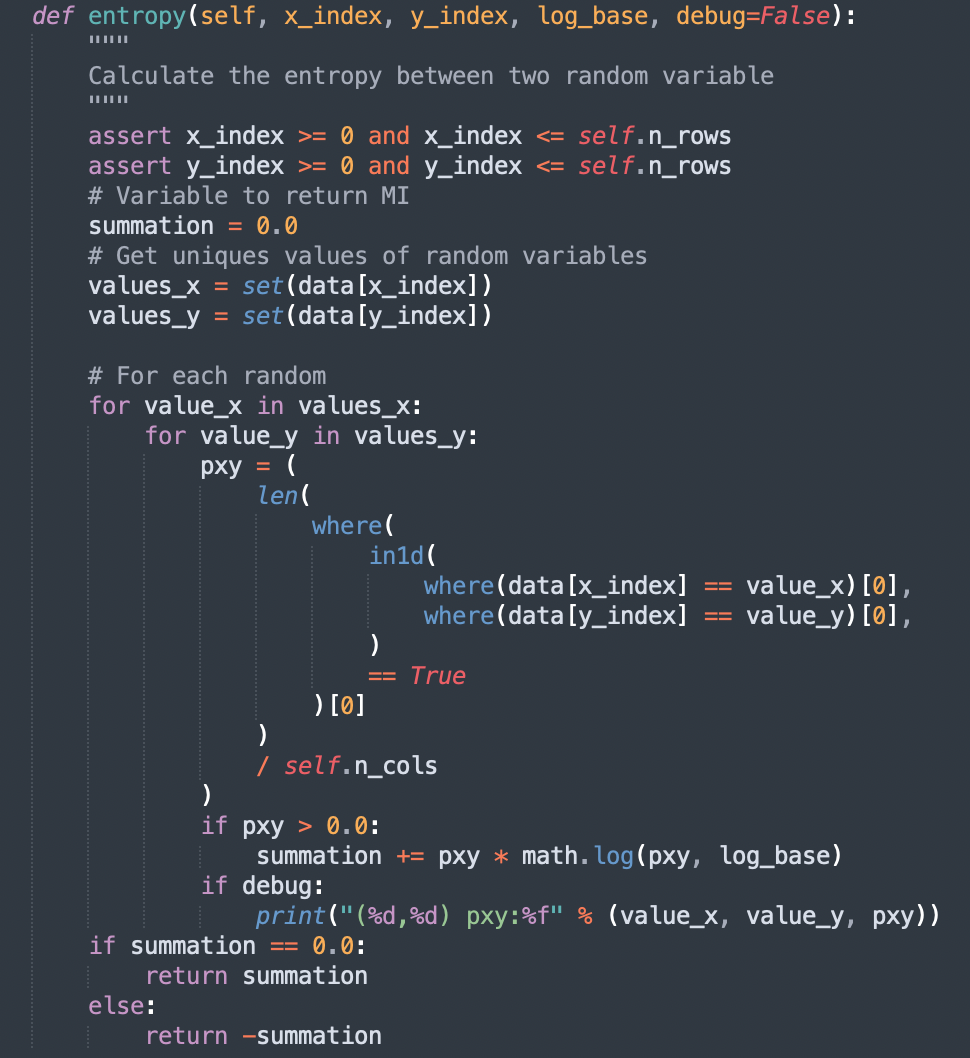


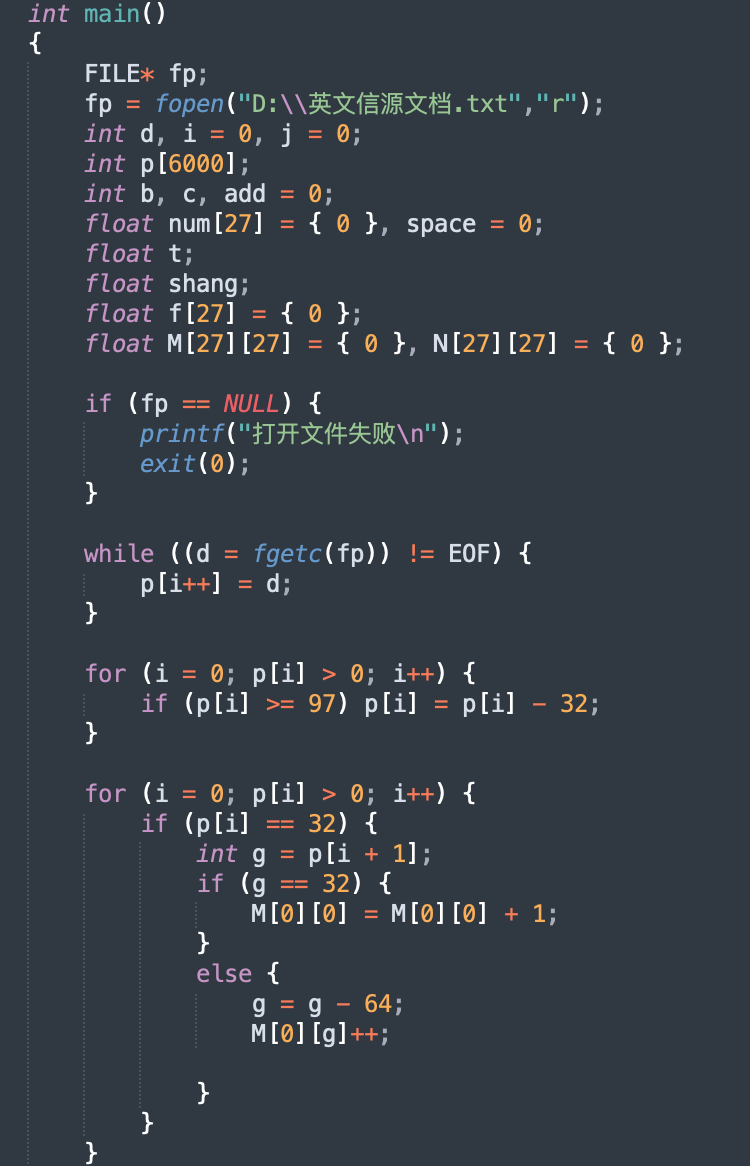
互信息：



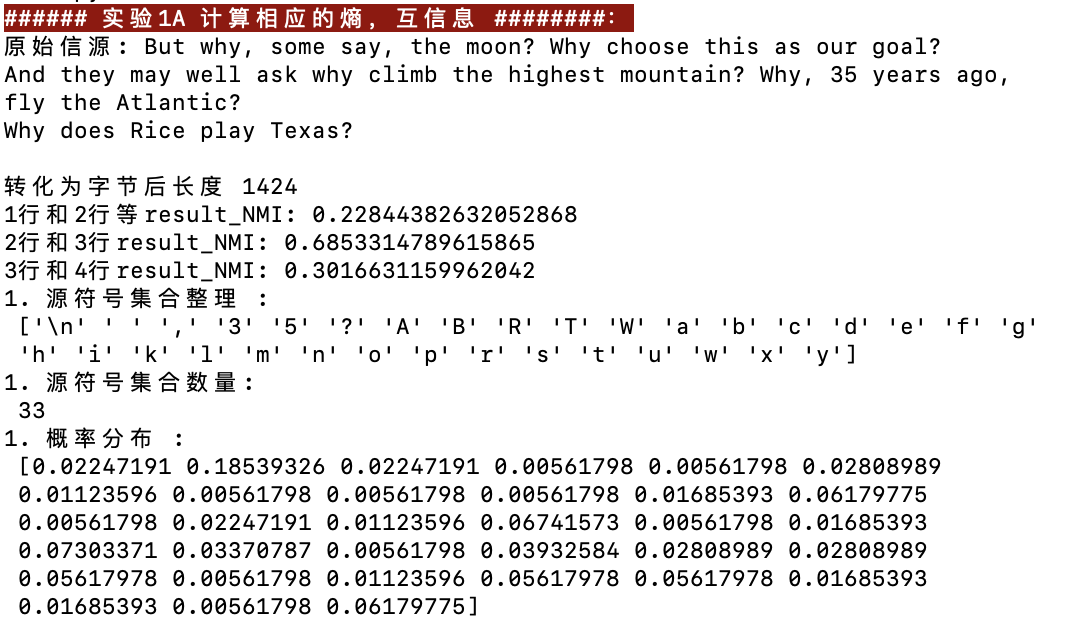
**实验2A：基于语料库，把文本信源依次看作是无记忆信源、一阶或二阶或更高阶马氏信源，计算相应的熵，平均互信息、互信息等**

## 程序代码（编程环境：python3/Dev-C++ ）

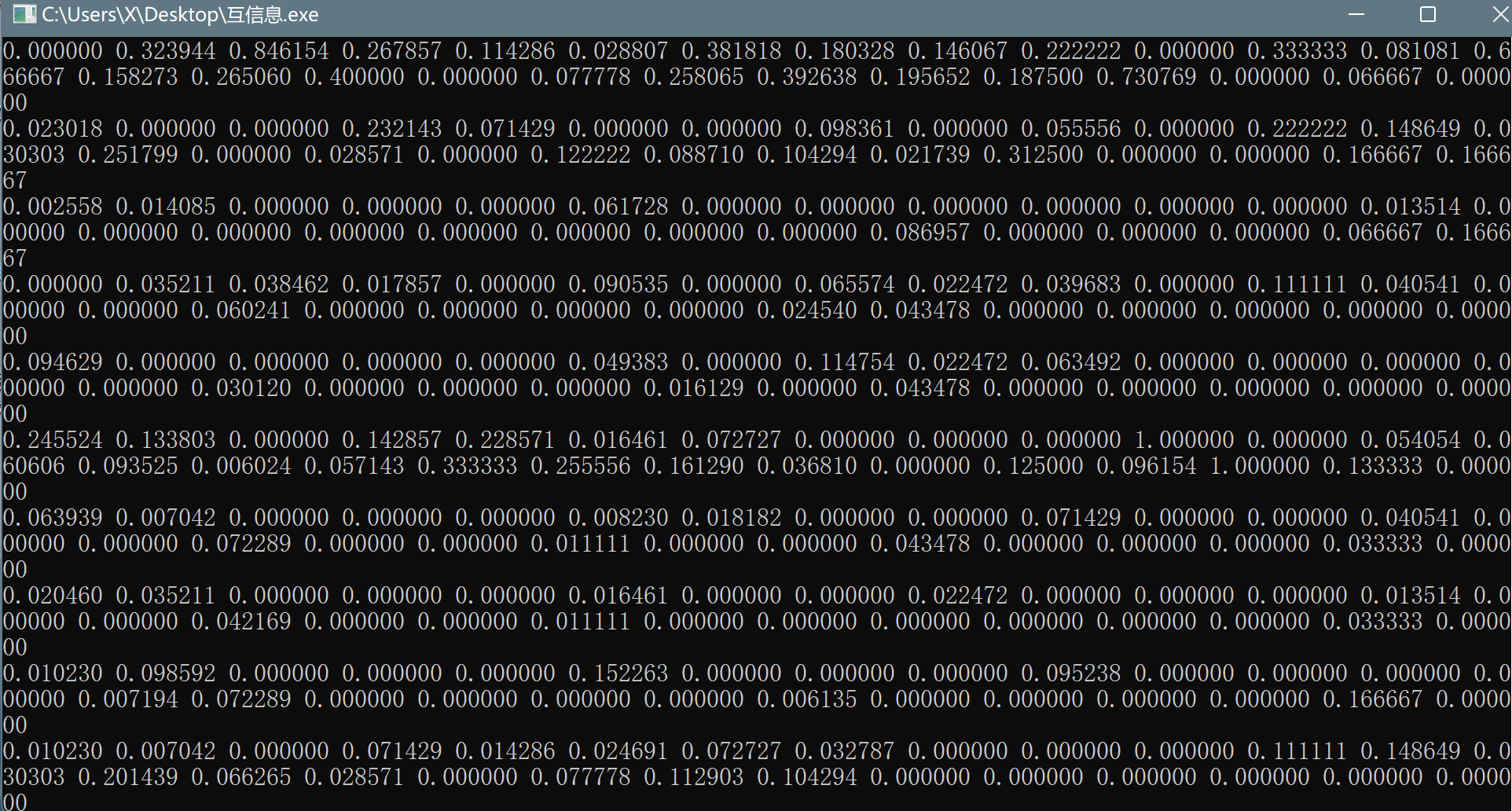




## 运行结果

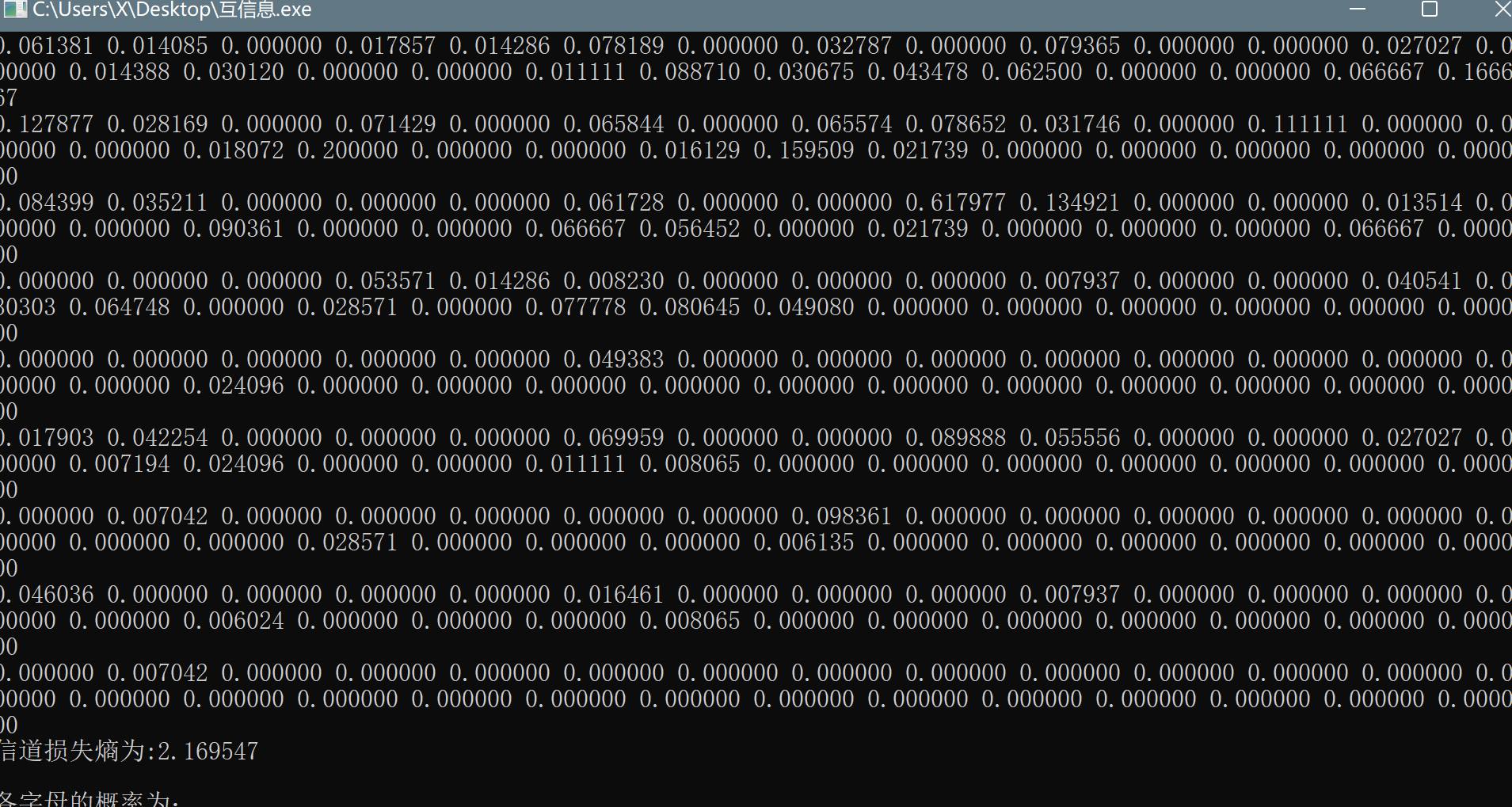
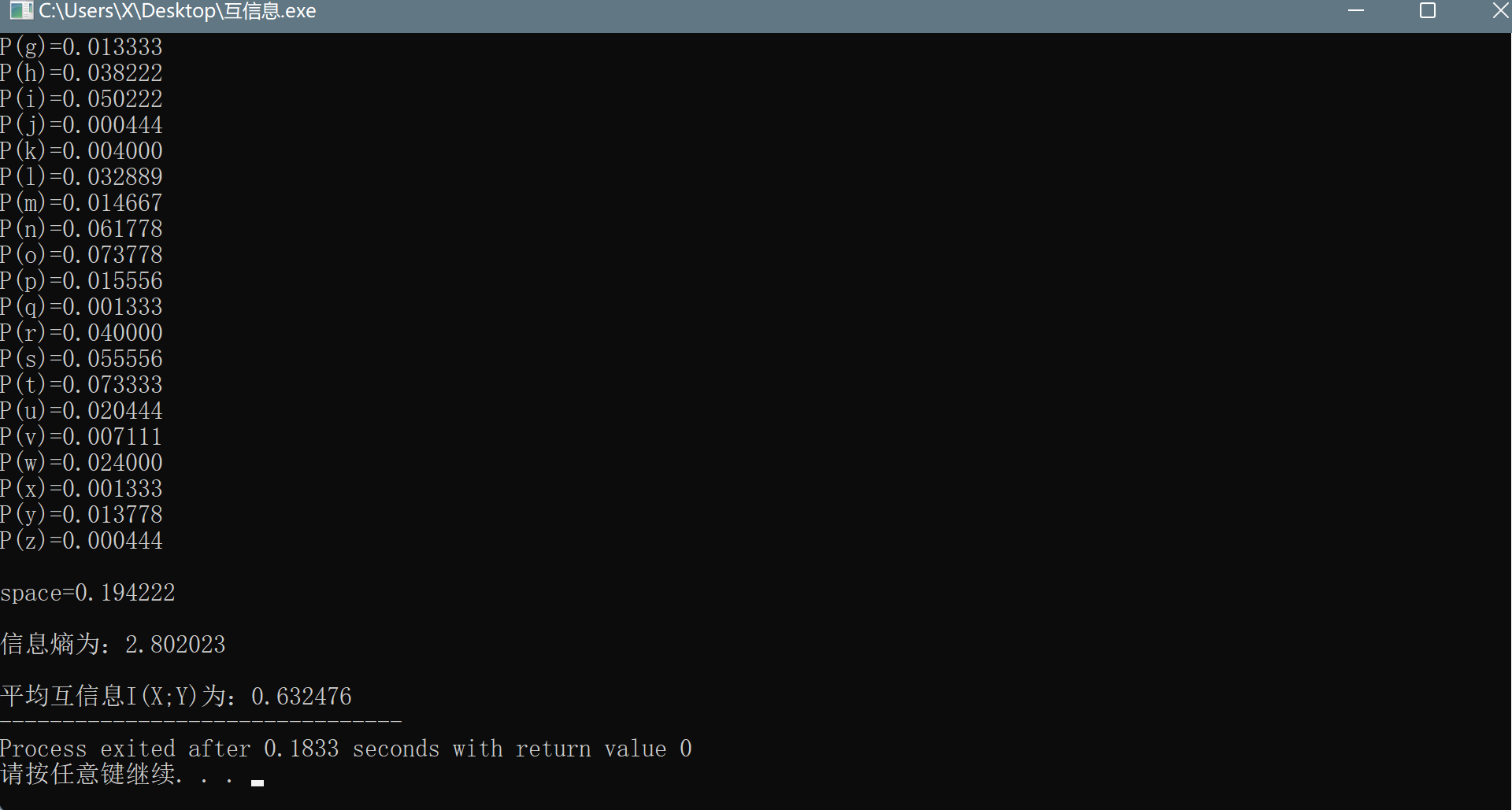


结果1：上图为平均互消息在句子之间的衡量



运行得到一个27\*27矩阵，代表每两个字母之间的互信息，发现矩阵中多数元素为0，即对应字母间互信息为0。

平均互信息为0.632476。

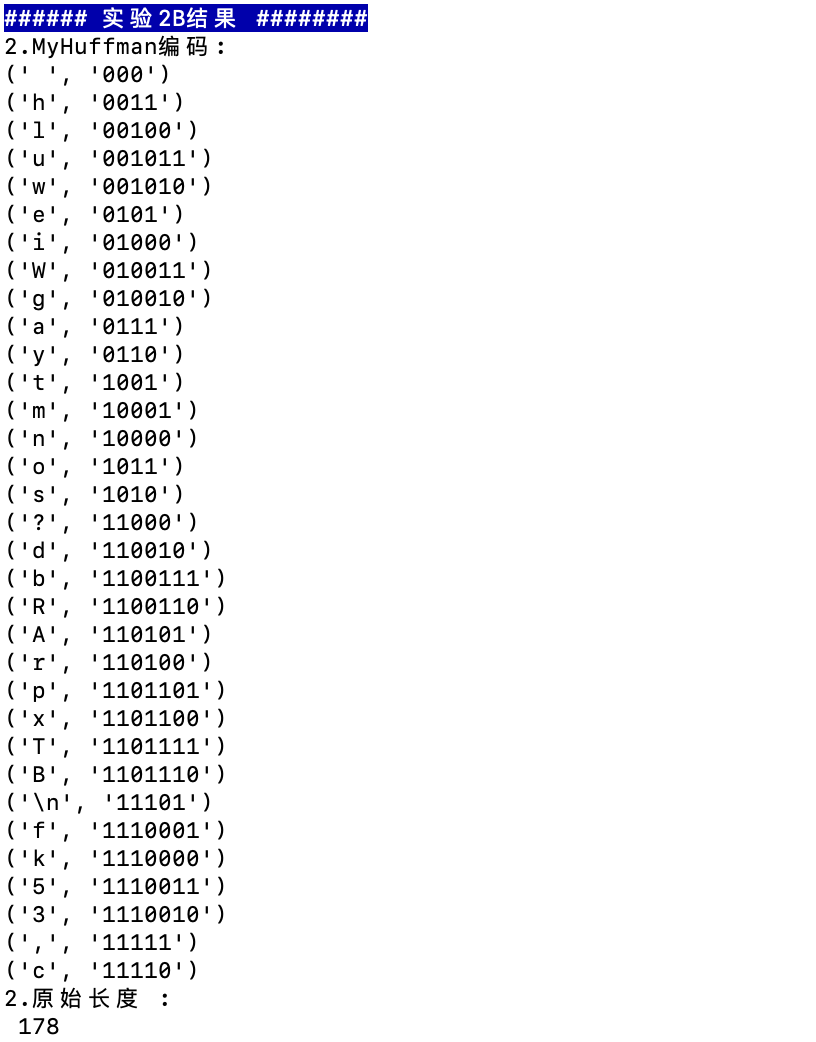
 

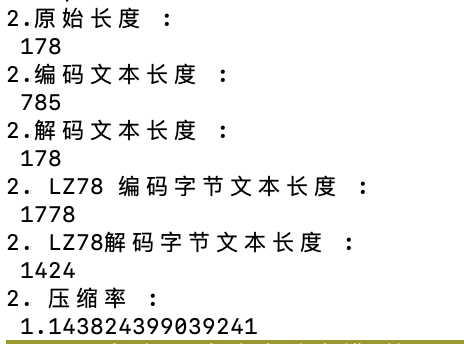
**实验2b：实现霍夫曼编码，对应使用LZ78算法，并比较压缩效果**

思路：为了实现霍夫曼树，我们选择将节点表示为字典，其键是树枝的标签，值是子树（节点）或叶子。 在编码字母表中选择键。 最后，树的叶子是源符号。

我们实现的第一步是为每个符号分配一个叶节点。 然后根据边缘概率将叶子插入优先级列。 优先级实际上被计算为概率的倒数，因此具有最低概率的符号具有最高优先级

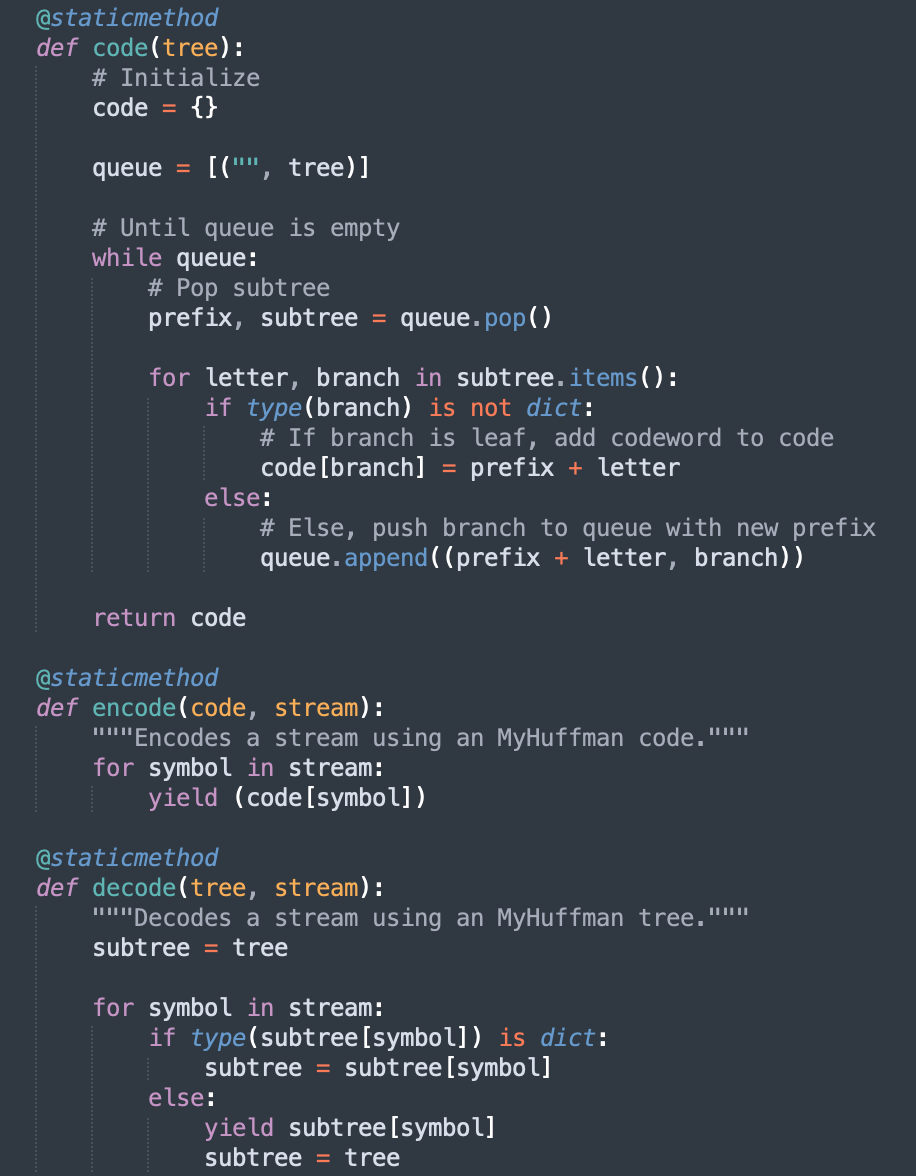
运行结果如下：

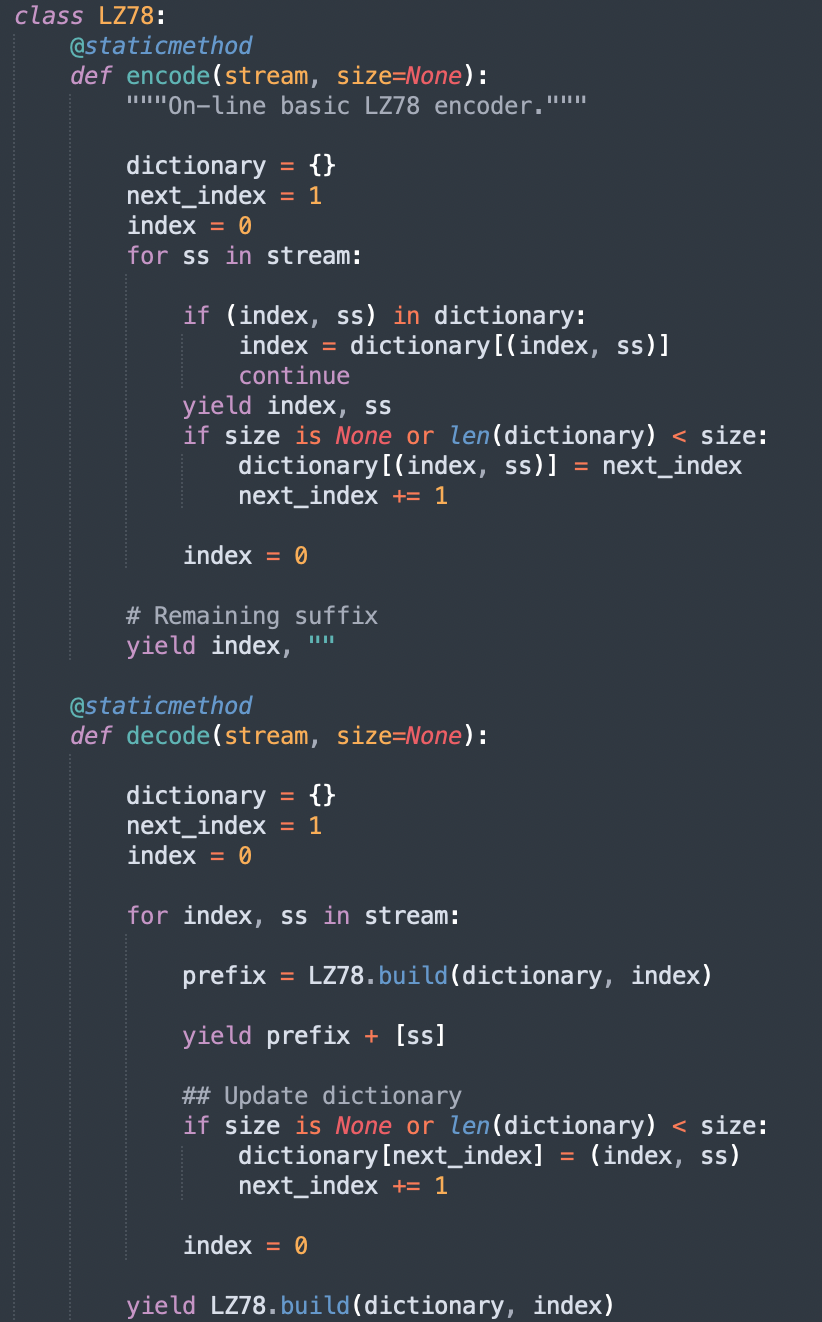




## 程序代码（编程环境：python3/Dev-C++ ）

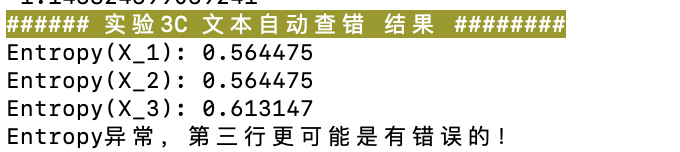






**实验2C：基于互信息等结果，实现文本自动查错**

阵结果如下：



## 程序代码（编程环境：python3/Dev-C++ ）

