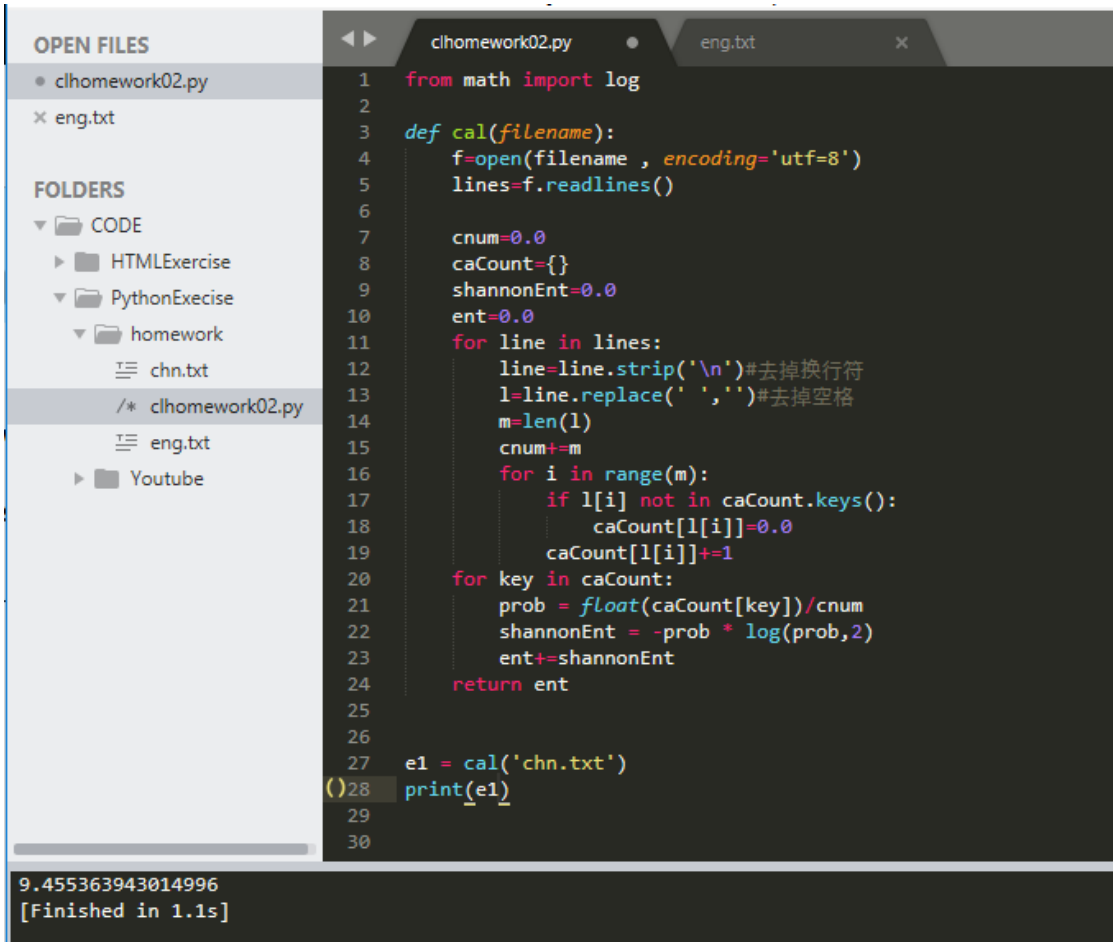
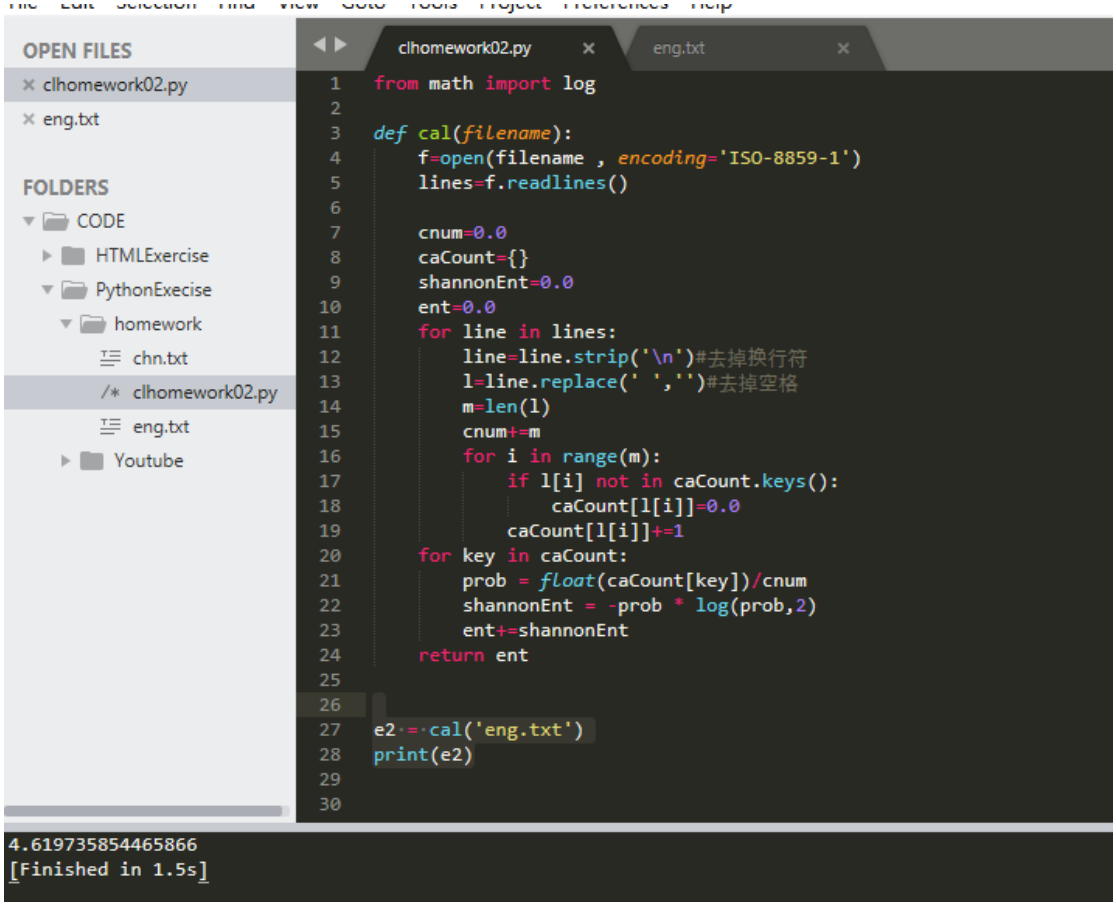
1.


OPEN FILES
clhomework02.py
eng.txt
FOLDERS
CODE
HTMLExercise
PythonExecise
homework
chn.txt
clhomework02.py
eng.txt
Youtube
clhomework02.py
eng.txt
from math import log
def cal(filename):
f=open(filename , encoding='utf=8')
lines=f.readlines()
cnum=0.0
caCount={}
shannonEnt=0.0
ent=0.0
for line in lines:
line=line.strip('\n')#去掉换行符
l=line.replace(' ','')#去掉空格
m=len(l)
cnum+=m
for i in range(m):
if l[i] not in caCount.keys():
caCount[l[i]]=0.0
caCount[l[i]]+=1
for key in caCount:
prob = float(caCount[key])/cnum
shannonEnt = -prob * log(prob,2)
ent+=shannonEnt
return ent
e1 = cal('chn.txt')
print(e1)
9.455363943014996
[Finished in 1.1s]



从这个方面看，平均信息熵越小，使用的比特数越少，这文字越好。但是事实并非如此。假设，当年中国的老祖宗创造中文时，仅发明两个文字“是”“不是”，那么中文的信息熵为 1 比特。是所有文字中最小的。但是这样好吗？

造成这样荒谬的结论的原因是并不是每个英文字母组成的词汇都是有用的。如”aa ,ab ,ac,…”所以，如果有人用汉字对比英文（在同样意义的词汇）的 byte 数，十有八九汉字要“节约”得多！

不容忽视的是中文的平均信息熵是 9.65 比特，在计算机信息作业的时候，汉字的每个字符需要两个字节的空间，因而中文的信息处理和传递的整体效率比英文等拼音文字的效率要低得多。尽管我们已经说明汉字实际上比英文和其他拼音文字只简不冗（从占用字节数的角度看），语言学上的问题仍然相当复杂，谁简谁繁似乎也还难以成为一种语言优劣的绝对定论。比如世界语、数学语言、电脑的汇编，显然都极简单而且规范，可是要代替自然的生活语言明显是不行的。因此，评价一种语言必须从多个方面考虑，仅考虑信息熵明显是不可行的。

2.

语言的输入随机变量 D，转化的汉字为 W。我想通过 W，了解 D。也就是用 P（D/W）来近似 P（D），现在定量地计算我们通过转化的汉字，了解语言的输入。

也就是用 P（D/W）来代替 P（D）编码，减少了多少不确定度

用 KL 距离来表征，就是：D（P（D/W）||P（D））。

接着，如果已知输入的语言 wi,可以选择最大 KL 距离 D（P（dj/wi）||P（dj））的 dj，也就是最可能汉字形式。

这就是 KL 距离的物理意义在实际中的一些运用。