README:

This is the first assignment for NLP course at UCAS in 2022.

This work which depend on a large scale of text is focus on calculating the entropy of English letter and Chinese letter.

I use the crawler tool to get the raw text data on the homepage of Xinhua net,the url is <https://english.news.cn/home.htm.>I get 3MB,6MB,9MB,12MB Chinese text data and 2.7MB English text data individually in order to calculate the probability of each character

appearing in the sample text and ultimately calculate the entropy.

The detail of the experiment is as follow:

1. Use the function:”get\_xinhua\_news” which is based on the crawler tool to grasp the content of the homepage and its linked url.
2. Use the function:”data\_clean”:to get rid of the punctuation and other useless symbol in the text. I recommend to take advantage of the coding section to find Chinese letter and English letter out. The division part of Chinese is ['\u4e00','\u9fa5'] while English is ['\u0061','\u007a'].
3. Use the function:”split” : to divide the whole text data into pieces, so that I can calculate the entropy with the increase text scale. In my experiment ,I set the interval to 3MB when dealing with the Chinese text.
4. Use the function:”calculate”: to calculate the entropy and list the character appearing in the text and list the probability of each character in the text.
5. Be careful that when the text size is 3000000,the total Chinese character is 1000000

The data and result are reserved in the txt file.