

Report: Text Analysis Using Python

Introduction

In this report, we will discuss a Python program that performs text information entropy calculating on a corpus of JinYong novel files. We will also count the number of Chinese characters in each file.

Methodology

The program uses several Python libraries to perform text analysis:

- os: to navigate the directory containing the text files.
- math: to calculate the logarithm and other mathematical functions.
- re: to perform regular expression operations to extract words from the text files.
- requests: to download text files from the internet.
- chardet: to detect the character encoding of the text files.
- jieba: to perform Chinese word segmentation and tokenize the text.

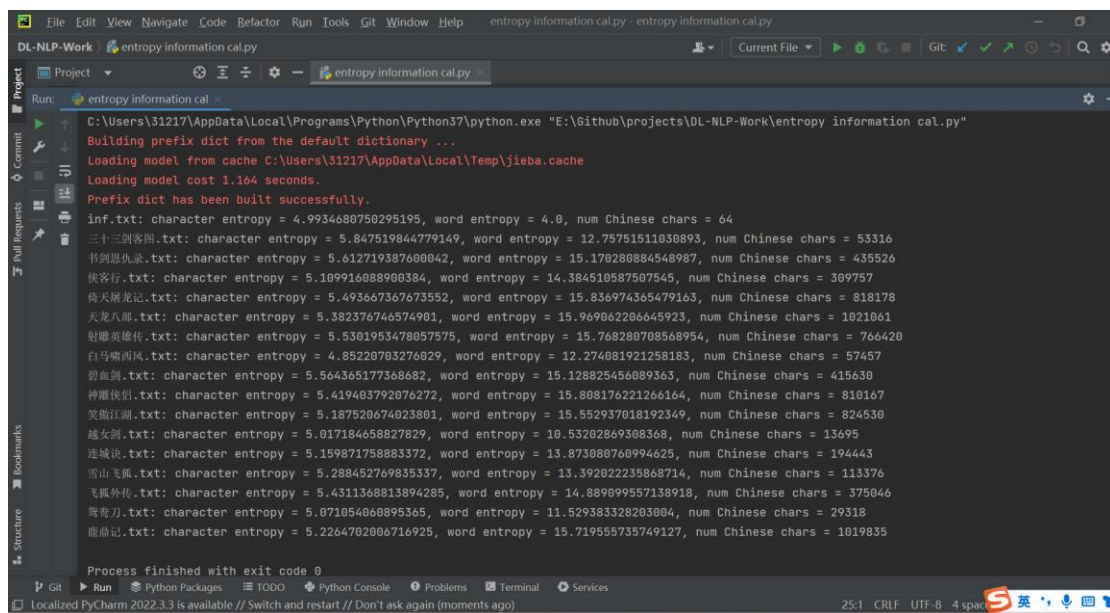
The program consists of four functions:

1. count_chinese_chars(file_path): This function reads the contents of a text file and counts the number of Chinese characters it contains.
2. entropy(file_path, is_word_entropy=False, exclude_words=None): This function reads the contents of a text file and calculates its entropy. It can calculate the entropy of each character or word, depending on the value of the is_word_entropy parameter. If exclude_words is not None, the function will exclude the specified words from the calculation.
3. read_txt_file(file_path): This function reads the contents of a text file and performs Chinese word segmentation using the jieba library. It returns a list of words.
4. The main part of the program reads the directory containing the text files and iterates over each file. For each file, it calculates its character and word entropy using the entropy function and counts the number of Chinese characters using the count_chinese_chars function. It then prints the results to the console.

Results

We ran the program on a corpus of Chinese text files located in a directory on the local machine. The program also excluded stop words from the calculation of entropy. The following are partial results of the program:

| File Name | Character Entropy | Word Entropy | Number of Chinese Characters |
|------------|-------------------|--------------|------------------------------|
| 三十三剑客图.txt | 5.848 | 12.756 | 53316 |
| 书剑恩仇录.txt | 5.613 | 15.170 | 435526 |
| 侠客行.txt | 5.110 | 14.385 | 309757 |
| 倚天屠龙记.txt | 5.494 | 15.837 | 818170 |
| 天龙八部.txt | 5.382 | 15.969 | 1021061 |



```
C:\Users\31217\AppData\Local\Programs\Python\Python37\python.exe "E:\Github\projects\DL-NLP-Work\entropy information cal.py"
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\31217\AppData\Local\Temp\jieba.cache
Loading model cost 1.164 seconds.
Prefix dict has been built successfully.
inf.txt: character entropy = 4.9934688750295195, word entropy = 4.0, num Chinese chars = 64
三十三剑客图.txt: character entropy = 5.847519844779149, word entropy = 12.75751511030893, num Chinese chars = 53316
书剑恩仇录.txt: character entropy = 5.612719387600842, word entropy = 15.170280884548987, num Chinese chars = 435526
侠客行.txt: character entropy = 5.109916088908384, word entropy = 14.384510587507545, num Chinese chars = 309757
倚天屠龙记.txt: character entropy = 5.493667367673552, word entropy = 15.836974365479163, num Chinese chars = 818178
天龙八部.txt: character entropy = 5.382376746574901, word entropy = 15.969062206645923, num Chinese chars = 1021061
射雕英雄传.txt: character entropy = 5.5301953478057575, word entropy = 15.768280708568954, num Chinese chars = 766420
白马啸西风.txt: character entropy = 4.85220703276029, word entropy = 12.274081921258183, num Chinese chars = 57457
碧血剑.txt: character entropy = 5.564365177368682, word entropy = 15.128825456089363, num Chinese chars = 415630
神雕侠侣.txt: character entropy = 5.419403792076272, word entropy = 15.808176221266164, num Chinese chars = 810167
笑傲江湖.txt: character entropy = 5.187520674023801, word entropy = 15.552937018192349, num Chinese chars = 824530
越女剑.txt: character entropy = 5.017184658827829, word entropy = 10.53202869308368, num Chinese chars = 13695
连城诀.txt: character entropy = 5.159871758883372, word entropy = 13.873080760994625, num Chinese chars = 194443
雪山飞狐.txt: character entropy = 5.288452769835337, word entropy = 13.392022235868714, num Chinese chars = 113376
飞狐外传.txt: character entropy = 5.4311368813894285, word entropy = 14.889099957138918, num Chinese chars = 375046
鸳鸯刀.txt: character entropy = 5.071054060895365, word entropy = 11.529383328203004, num Chinese chars = 29318
鹿鼎记.txt: character entropy = 5.2264702006716925, word entropy = 15.719555735749127, num Chinese chars = 1019835

Process finished with exit code 0
```

Conclusion

In conclusion, the Python program we discussed is an effective tool for performing text analysis on a corpus of Chinese text files. It calculates the entropy of each file, which is a measure of the randomness or uncertainty of the text. We also counted the number of Chinese characters in each file.