Report: Information Entropy Calculation

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.

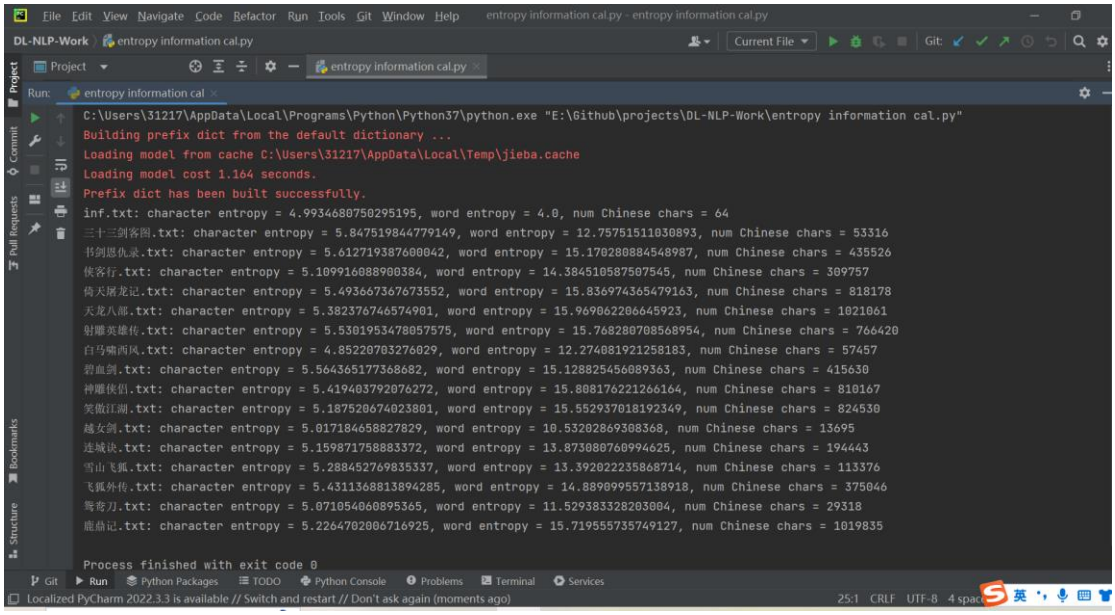
The program consists of four functions:

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



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.