After data cleaning, we re-generate the line figure for the number of reported results, //

According to the observation and analysis of the new line figure, the data fluctuate greatly in the early period, and gradually tend to be stable and show a slow decline in the later period. Generally speaking, the data of the number of reported results is still in a state of fluctuation over time, which belong to the non-stationary time series. Based on the above analysis, we decide to build the ARIMA model to solve the problem.

In order to study the relationship between word attributes and the proportion of difficult patterns in report scores, after searching and learning relevant materials, we define word attributes and divided them into 3 criteria, respectively

1) Letter frequency, that is, the proportion of the number of times that each letter in the word appears in the sum of all letters in the word data to the total number of letters. The calculation formula is

(For example, if the letter "e" appears 15 times in a total of 25 words, its letter frequency is.)

2) Frequency of combined sounds, similar to letter frequency, that is, the proportion of the number of combined sounds in all word data contained in this word to the total number of combined sounds, the calculation formula is

3) Word frequency. This data mainly comes from walfram, a professional word frequency statistics website, which records the usage frequency of each word worldwide. Since there is a large gap between some words, it is not good for comparison, we use sigmoid function (The functional relation is) for data processing. Its domain of definition is limited to. The processed values are sorted, and the value of the 180th word in the word frequency sorting of 359 words is processed as . The image of the specific result is shown below