COMP 90049 report

Knowledge Technology Project

Analysis Approximate matching methods

Student ID: 792715

Student Name: Junwen Xie

1. Introduction

The purpose of this project is to programme a single application with multiple approximate matching methods which can identify the film title in these reviews, and also, the application can evaluate the quality of the film through these IMDb reviews.

The aim of this project is to write a simple Python programme include N-gram, Local Edit distance and Global Edit distance methods to calculate the possible film title for these film reviews. The application will read both film title and review data from txt file and put these data into one of the approximate matching methods. The programme will automatically figure out a possible film title for the reviews. This report will analysis the basic problems of these approximate methods and show the possible solution for these problems.

2. Approximate matching methods

When using the function N-gram, the methods will set these strings in a vector space which allowing the sequence to be compared with each other. The lower N-gram distance the programme gets the better matching with the film title.

Edit distance is a way to calculate the two sting dissimilarly. The advantage of the edit distance methods which is needed to compare all the single strings to each other. The disadvantage of the methods is the process speed too slow and the final result is not accuracy.

Problem analysis:

Compare to the Edit distance methods, the advantage of N-gram method has better effectiveness than any others. The N-gram function can get the final result faster. However, the N-Gram may have difficulties to figure out numerous articles. The following table shows the performance of the N-Gram when processing the approximate matching.

N-Gram	**Film Title example**	**Reviews' wordCount**	**Performance**
2-Gram	[ab], [bc], [cd]…	150	Normal
		800	Inaccurate
		More than 1500	Bad

N-Gram	**Film Title example**	**Reviews' wordCount**	**Performance**
3-Gram	[abc], [bcd], [cde]…	150	good
		800	Normal
		More than 1500	Inaccurate

N-Gram	**Film Title example**	**Reviews' wordCount**	**Performance**
4-Gram	[abcd], [bcde], [cdef]…	150	good
		800	Normal
		More than 1500	Inaccurate

The main reason for the N-gram performance change is the length of the film titles. The working principle of the N-gram is to calculate the matching distance between film titles and film review. The problem of the matching method which is the longer film title has possible matching strings in the review. Therefore, the longer film titles always have the advantages of the result.

Let's take an example: searching with Film Title [`Yor, the Hunter from the Future`] and Film Title [The Hunter] in one of the same review. The result of the N-Gram distance will prefer the longer one, because of the more strings they have, the more possibilities for these strings match with each other. This will be very difficult for the N-gram to decide which one is the right film titles for the film review.

Another problem of the N-gram is the number of 'N'. The N-gram function is using the length of 'N' to divide the string in a vector space. 'N' is the length of each matching part. However, some of the film titles are smaller than the 'N' length such as film title [9], film title [15]. If the 'N' is set up as 3, the string cannot match any other string because the title is too short. It will cause the miss compare of the right title to the film reviews.

The Edit distance methods are also influenced by the length of the strings. After using the precision, the formula to test the precision. [P = TP/(TP+FP)] the value of precision is only about 9.5% which is a lower score to search the film titles correctly.

Problem-solving:

To decrease the deviation of theses N-gram problems. The N-gram distance score of the film title matching should be treated fairly due to the different length of the film titles. One of the methods is using the formula: ***p = N-Gram Distance Score / Length of the film title***. This simple formula can help the result become a relative value by comparing with the length of the film title.

Another solution to improve the accurate of the string matching which is changing the value 'N' of N-gram according to the length of the review. If a review has the long paragraph, the value 'N' will be set up a large number. However, some length of the film titles is smaller than the length of 'N'. It is better to use exact search rather than approximate matching.

Due to the reason above, the result of the N-gram and Edit Distance approximate matching method for the review is still not too good to reach to expectation.

3. Film Quality Evaluation methods

While reading a film review, there must be some positive or negative evaluation of these films. The quality evaluation method is a function that can help the program to draw a clear distinguish between good and bad.

To reach to this project, it is very important to create a database for some words which are usually used by making a comment such as:

some positive word like:
[*'good','cool','nice','exciting','awesome','love','like','perfect','interesting','lovely'…*]
some negative words like:
[*'boring','dislike','bad','terrible…'*]

This kind of words always appears in the film reviews. After creating the word evaluation database. We can use the approximate matching methods such as N-gram to match these strings to the review list. Therefore, the function will get two different values for both positive and negative matching distance. We can now compare the result: if the positive value larger than the negative value, we can assume that this review is a positive review.

However, the approximate matching methods have an average error while doing the matching process. The better methods for calculating the quality evaluation is the exact search methods. And also improving the wide range of the lexicon of the quality evaluation can also help provide the precision of the final result.

4. Conclusions

In conclusion, both of the N-gram methods and the edit distance methods have the ability to identify the film titles in these reviews. However, these approximate matching methods have difficulty in identifying the longer paragraph review which includes more strings. The accuracy will be reduced due to these problems. Although by trying some of the precision methods to provide a fairly result the approximate still have wide margins to identify the film titles from these reviews. It is better to use multiple approximate methods to test the value of the result and provide more effective values by mixing the analysis of the result.

References

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

Zobel, Justin and Philip Dart. (1996). Phonetic String Matching: Lessons from Information Retrieval. In Proceedings of the Eighteenth International ACM SIGIR Conference on Research and Development in Information Retrieval. Z¨urich, Switzerland. pp. 166–173.