Term project report

소프트웨어학부

20186663

전찬웅

Main idea

I thought that there were many words of the same kind or similar to the subject in the subject problem. So I decided fit the problem by using topic modeling like latent Dirichlet allocation or cosine similarity.

Finally, my project's main idea is "Fit the English topic, subject problems in the Korean SAT with TF-IDF Vectorization or Count Vectorization and using cosine similarity.".

System development

First, I used the past Korean SAT English part problems. And then I saved these problems as csv files. Csv file contains documents, sentences, answers, respectively.

Second, in the program code, after vectorizing each problem largely, I found the cosine similarity between the document and the sentences, and then decided the correct value with the largest cosine similarity value. In the vectorizing each problem, I used Count Vectorization and TF-IDF Vectorization.

Evaluation

Before the project, I tried to evaluate through the f1 score, but the problem was five multiple choices, so I couldn't apply the f1 score. Therefore, the

distribution of the answers to the problem is almost the same, and since accuracy is used as an evaluation indicator when solving a real problem, accuracy is used as an evaluation indicator.

Conclusion

The results evaluated through the above evaluation indicators are as follows.

Count Vectorization : `accuarcy 0.4925373134328358`

TF-IDF Vectorization : `accuarcy 0.47761194029850745`

Count Vectorization & LDA : `accuarcy 0.26865671641791045`

TF-IDF Vectorization & LDA : `accuarcy 0.3880597014925373`

According to the above results, the accuracy of count vectorization was the highest, and overall, the value of not doing LDA was better.

When considering the above results, it was difficult to find a meaningful one in each sentence because the amount of each sentence was too small length compared to the document. Therefore, the value of count vectorization, which simply vectorizes into the number of words, came out the highest, and in short sentences, the meaning of the sentence being LDA was difficult to find, so the process of LDA was hindered.

Reference

I got a lot of information from Google search, such as vectorization, cosine similarity, and Latent Dirichlet allocation and so on.

I used python language for programming.

I used pandas, numpy, sklearn libraries for data analysis.