11-791 Homework 4

Wenyi Wang (Andrew ID: wenyiwan)

1. Building Vector space Retrieval Model using UIMA

1.1 Design

With the program template provided in the homework, it is easy to fill in all the blanks. I hardly did any modification to the type system, the way of generating type systems using UIMA>>JCasGen is similar to that in previous homework (task 0). Basically, what I modified are the class annotators and class retrieval evaluators. Details are as follows.

1.1.1 Annotators

Here I extracted bag of words feature vectors from the input text collection (task 1). In addition, I use a dictionary to store the stop words in the file provided in the homework. In this way, most stop words in the input can be ignored in order to increase accuracy. As for tokenization, I don't use external packages and make a simple implementation from scratch. The non-alphabetical symbols are removed in my design.

1.1.2 Retrieval evaluators

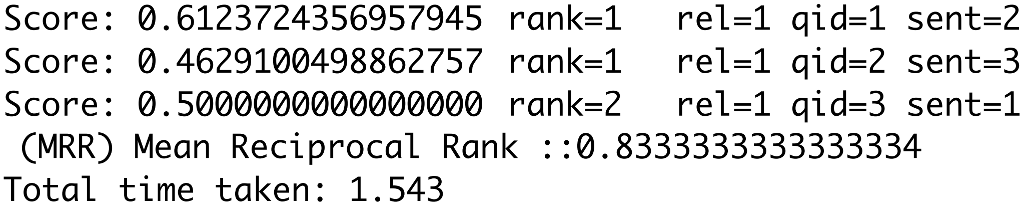
Basically, I use lists to store the similarity score and rank of each answer.

Regarding input, a general method is used to read queries and answers. Although the field “rel” indicates correctness and distinguished answers and queries, it may not be a generic representation. So I decide to check “qid” every line to tell whether the line is a query or answer.

Regarding similarity computation, there are three methods implemented in this class, namely, cosine similarity (task 2), Dice coefficient and Jaccard coefficient (task 4). MRR computation is also implemented in this class, as indicated by the template (task 3).

1.2 Result

The result of the original implementation of cosine similarity is as follows.



```
Score: 0.6123724356957945 rank=1   rel=1 qid=1 sent=2
Score: 0.4629100498862757 rank=1   rel=1 qid=2 sent=3
Score: 0.5000000000000000 rank=2   rel=1 qid=3 sent=1
 (MRR) Mean Reciprocal Rank ::0.8333333333333334
Total time taken: 1.543
```

2. Error Analysis

From the result shown above, there is an error in query #3. Instead of the first answer, the system gives the highest score to the second answer. I noticed that the second answer is much longer than the other two answers. The difference between the lengths of two sentences may affect the score computed by cosine similarity.

I tried Dice coefficient, Jaccard coefficient, as suggested in the homework. The result didn't change much. I also tried angular similarity, which is a modification of cosine similarity and computed as follows:

$$1 - \left(\frac{\cos^{-1}(\text{cosine\_similarity})}{\pi}\right)$$

The result didn't change though.

After discussion with my fellow students, I tried to do modification to the lengths of sentences and tried to change the power of the denominator in the original computation of cosine similarity. Based on my trials, when the power equals 1.5, 1.6 or 1.7 (instead of 2 originally), the result is as follows:

```
Score: 0.2013453940822393 rank=1   rel=1 qid=1 sent=2
Score: 0.1251293987808771 rank=1   rel=1 qid=2 sent=3
Score: 0.1894645708137998 rank=1   rel=1 qid=3 sent=1
 (MRR) Mean Reciprocal Rank ::1.0
Total time taken: 1.601
```

The new MRR reaches 1.0, which proves that the lengths matters and that my technique is effective. In hindsight, I think the modification works because reducing the denominator makes the numerator more significant, thus amplifying the difference between the two vectors.