Report for 11791 HW4

Jingwei Shen

js1

1. Overview

In this assignment, I first implement the retrieval system with cosine similarity as required. Then I make improvements using dice coefficient. Based on the comparison I think the weaknesses in the original retrieval system lie in the score function (cosine similarity) and stopword.

2. Implementation with cosine similarity

The basic idea behind this method is to calculate the similarity between two sentences based on the frequency of key words in each. Key words are selected by eliminating the irrelevant words called stop words. Similar to calculating the cosine of the angle of two vectors, the cosine similarity is defined as the cosine between two word vectors

Here is the result of my implementation.

```
score: 0.5163977794943222 rank= 1 rel = 1 qid=1 Classical music is dying
score: 0.16666666666666666 rank= 1 rel = 1 qid=2 Energy plays an important role in climate change
score: 0.5 rank= 2 rel = 1 qid=3 One's best friend is oneself
score: 0.18257418583505536 rank= 2 rel = 1 qid=4 The shortest distance between new friends is a smile
score: 0.0 rank= 2 rel = 1 qid=5 It takes a long time to grow an old friend
 (MRR) Mean Reciprocal Rank ::0.7
Total time taken: 0.741
```

We see that there are still some errors existing, leading to a MMR less than 1.

3. Implementation with dice coefficient

Dice coefficient is another method to measure the similarity between two sentences. After extracting the key words, dice coefficient is defined as the ratio of the size of their intersection and the size of their union. Here is the result after I change the new score function with dice coefficient.

```
score: 0.3333333333333333 rank= 1 rel = 1 qid=1 Classical music is dying
score: 0.3333333333333333 rank= 1 rel = 1 qid=2 Energy plays an important role in climate change
score: 0.3333333333333333 rank= 1 rel = 1 qid=3 One's best friend is oneself
score: 0.1 rank= 2 rel = 1 qid=4 The shortest distance between new friends is a smile
score: 0.125 rank= 1 rel = 1 qid=5 It takes a long time to grow an old friend
 (MRR) Mean Reciprocal Rank ::0.9
Total time taken: 0.859
```

4. Error analysis

I think error mainly comes from the score function. Cosine method cares more about the frequency of key words, it will work well in most cases. However, if the two sentences use metaphor, then the cosine value might probably be low.

For example :
S1 : If you run after two hats, you will catch neither.
S2: One should work with all his heart.

We see that the above two sentences actually share the same meaning, however their cosine value is fairly low .

Also the stop word list is important. If we eliminate all the words like“no, not”, then two sentences might have opposite meanings even if the cosine is high.

For example :
S1: John loves Marry.
S2: John doesn’t love Marry.

These sentences share a high cosine while have the opposite meanings.