CARNEGIE MELLON UNIVERSITY

Engineering and Error Analysis with UIMA

Mengda Yang

Andrew ID: mengday

October 26, 2013

1 STRAIGHT FORWARD APPROACH

The task of this homework is to fill in to the blanks of the document, finish the program and get the result. By modifying the least code I get the following result.

```
Origional
Score:0.452267 rank=1 rel=1 qid=1 Classical music may never be the most popular
music
Score:0.102062 rank=1 rel=1 qid=2 Climate change and energy use are two sides
of the same coin.
Score:0.507093 rank=1 rel=1 qid=3 The best mirror is an old friend
Score:0.172133 rank=3 rel=1 qid=4 If you see a friend without a smile, give
him one of yours
Score:0.000000 rank=3 rel=1 qid=5 Old friends are best
(MRR) Mean Reciprocal Rank ::0.7333333333333334
Total time taken:  0.337
```

2 LOWER CASE AND PRECISE WORD EXTRACTION

Since we are primarily using the bag-of-word schema, one trivial improvement to do is to change all the letters into lower case. By applying this, I get the following result.

```
To lower case
Score:0.452267 rank=1 rel=1 qid=1 Classical music may never be the most popular
music
Score:0.306186 rank=1 rel=1 qid=2 Climate change and energy use are two sides
of the same coin.
Score:0.507093 rank=1 rel=1 qid=3 The best mirror is an old friend
Score:0.172133 rank=3 rel=1 qid=4 If you see a friend without a smile, give
him one of yours
Score:0.158114 rank=1 rel=1 qid=5 Old friends are best
(MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken:  0.334
```

We can see there is a significant improvement on the 5th query, that's because the *Old* with a capital *O* in the candidate sentence did not appear in the query sentence.

We also want to get rid of the non-alphabetic characters at the end of each word, so after applying this technique, we can get

```
Trim non-alphabet symbols after words
Score:0.452267 rank=1 rel=1 qid=1 Classical music may never be the most popular
music
Score:0.306186 rank=1 rel=1 qid=2 Climate change and energy use are two sides
of the same coin.
Score:0.507093 rank=1 rel=1 qid=3 The best mirror is an old friend
Score:0.258199 rank=3 rel=1 qid=4 If you see a friend without a smile, give
him one of yours
Score:0.158114 rank=1 rel=1 qid=5 Old friends are best
(MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken:  0.337
```

3 DELETE STOP WORDS

There are multiple ways to delete stop words. I can either choose to delete them with Stanford NLP toolkit with their tokenizers, or I can just use the stop word dictionary provided in the archetype. I chose the latter. I created a new class `StopwordFilter` inside the class `Utils` to utilize the algorithm and got the following result.

```
Delete stop words
Score:0.612372 rank=1 rel=1 qid=1 Classical music may never be the most popular
music
Score:0.462910 rank=1 rel=1 qid=2 Climate change and energy use are two sides
of the same coin.
Score:0.500000 rank=2 rel=1 qid=3 The best mirror is an old friend
Score:0.182574 rank=2 rel=1 qid=4 If you see a friend without a smile, give
him one of yours
Score:0.235702 rank=1 rel=1 qid=5 Old friends are best
(MRR) Mean Reciprocal Rank ::0.8
Total time taken:  0.406
```

In this result we can see a performance (MRR) drop, the ranking of two of the correct candidates changed. By looking at the 3th query, we can see that before the stop word deletion, the correct candidate wins solely because that it has a shorter sentence, while after the deletion a false candidate wins because the word *best* appeared in it twice. This may lead to a conclusion that to find out the true answer, we need some more advanced algorithms than the bag-of-word schema.