HW4 Report

NAME: Ying Sheng
ID: yingshen

[We expect you to document how you designed the execution and deployment architecture for the sample information system]

DocumentReader: create doc annotation contains: text, QueryID and relevance value.
DocumentVectorAnnotator:

1. construct the tokenList from the text in doc annotation
    a. for each distinct word in doc.getText() create a Token annotation with <String word, Integer count> pair.
    b. Construct an FSList use the Tokens got above
2. add tokenList doc annotation.

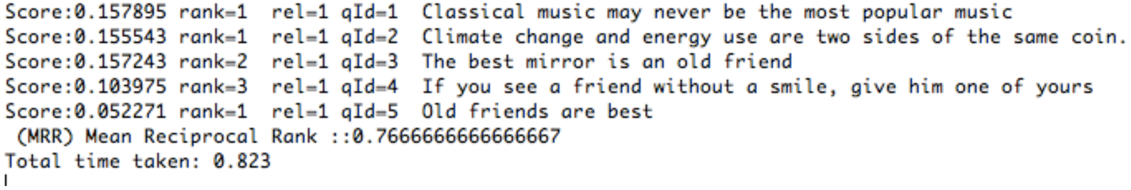
RetrievalEvaluator: contains 2 main function

1. processCas: create a global dictionary and the word freq. vector for every sentence

   Since processCas process only one document annotation for each iteration, new public variable HashSet<String> library and ArrayList<HashMap<String, Integer>> are created to store the result.

2. collectionProcessComplete:
    a. Use the dictionary to fill in word freq. vector, such that each vector has the same size.
    b. Calculate cosine-similarity score for each retrieval
    c. Get the rank for correct retrieval
    d. Calculate MRR and output to stout.

Error Analysis:
Here's the result for the extended input data (with 5 questions)

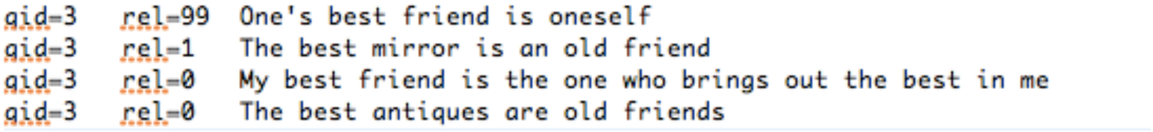


```
Score:0.157895 rank=1  rel=1 qId=1  Classical music may never be the most popular music
Score:0.155543 rank=1  rel=1 qId=2  Climate change and energy use are two sides of the same coin.
Score:0.157243 rank=2  rel=1 qId=3  The best mirror is an old friend
Score:0.103975 rank=3  rel=1 qId=4  If you see a friend without a smile, give him one of yours
Score:0.052271 rank=1  rel=1 qId=5  Old friends are best
 (MRR) Mean Reciprocal Rank ::0.7666666666666667
Total time taken: 0.823
```

The error part is the qID=3 and qID=4, which ranks 2 and 3 instead of 1 in the cosine similarity ranking.

For qID=3:

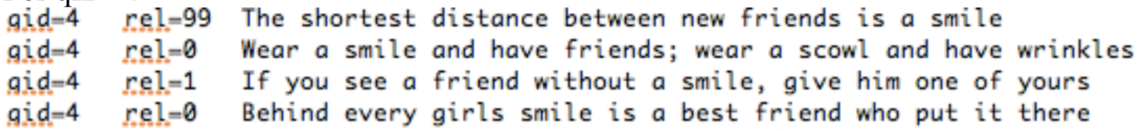
.



```
qid=3   rel=99  One's best friend is oneself
qid=3   rel=1   The best mirror is an old friend
qid=3   rel=0   My best friend is the one who brings out the best in me
qid=3   rel=0   The best antiques are old friends
```

The retrieval in the middle “My best friend is the one who brings out the best in me” has the highest cosine similarity score. The reason is: the correct answer use metaphor ‘mirror’, which makes it didn’t work well when only using the token without considering any meaning or relation of the word.

For qID=4



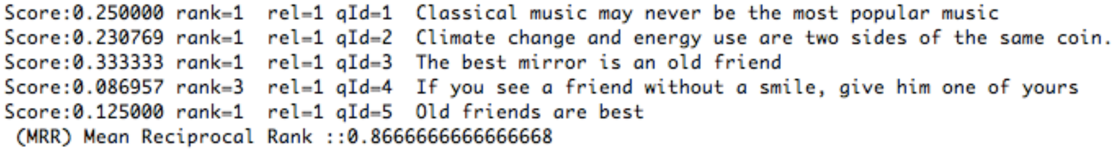
```
qid=4   rel=99  The shortest distance between new friends is a smile
qid=4   rel=0   Wear a smile and have friends; wear a scowl and have wrinkles
qid=4   rel=1   If you see a friend without a smile, give him one of yours
qid=4   rel=0   Behind every girls smile is a best friend who put it there
```

The correct answer ranks 3 because: 1) friend and friends don’t match in our method 2) “give him one of yours”, the token ‘one’ means the same thing as smile, we don’t have such information in our method.

BONUS PART:

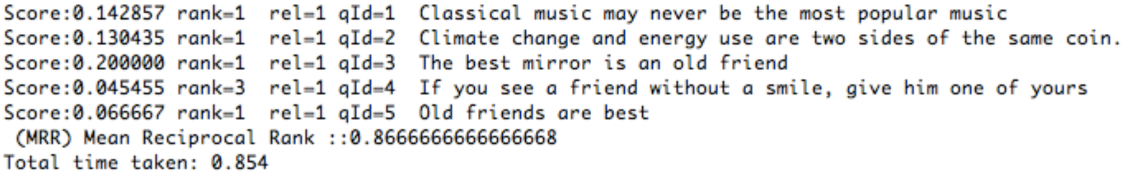
I tried both dice coefficient and Jaccard coefficient, here’s the results:

Dice:



```
Score:0.250000 rank=1  rel=1 qId=1  Classical music may never be the most popular music
Score:0.230769 rank=1  rel=1 qId=2  Climate change and energy use are two sides of the same coin.
Score:0.333333 rank=1  rel=1 qId=3  The best mirror is an old friend
Score:0.086957 rank=3  rel=1 qId=4  If you see a friend without a smile, give him one of yours
Score:0.125000 rank=1  rel=1 qId=5  Old friends are best
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
```

Jaccard:



```
Score:0.142857 rank=1  rel=1 qId=1  Classical music may never be the most popular music
Score:0.130435 rank=1  rel=1 qId=2  Climate change and energy use are two sides of the same coin.
Score:0.200000 rank=1  rel=1 qId=3  The best mirror is an old friend
Score:0.045455 rank=3  rel=1 qId=4  If you see a friend without a smile, give him one of yours
Score:0.066667 rank=1  rel=1 qId=5  Old friends are best
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken: 0.854
```

From the result, we can see the result is not slightly improved. The MRR in dice and jaccard ranking is 0.867, a little better than the cosine similarity ranking’s MRR 0.767