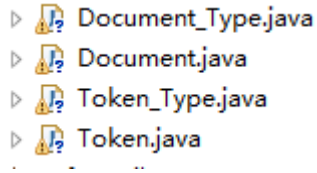
10791 hw4report
Engineering and Error Analysis with UIMA

Andrew ID: jiacongh
Finished Date: 10/28/2013

In this task, I finished the following tasks.
0. automatically generate the type system implementation using UIMA>>JCasGen
1. correctly extract bag of words feature vector from the input text collection
2. compute the cosine similarity between two sentences in the text collection
3. compute the Mean Reciprocal Rank (metric) for all sentences in the text collection
4. Design a better retrieval system that improves the MRR performance measure

1. Generate the type system

In this task, I used the type system supplied to finish the work. The type system is as follow.



2. Extract bag of words

In this task, I extracted bag of words and compute each words' frequency in its sentence. Then I use a token object to express each word and use setFrequency() method to put the frequency to each word.

```
docText:  Pop music has absorbed influences from most other genres of popular music
token:pop       frequency:1
token:music     frequency:2
token:has       frequency:1
token:absorbed  frequency:1
token:influences        frequency:1
token:from      frequency:1
token:most      frequency:1
token:other     frequency:1
token:genres    frequency:1
token:of        frequency:1
token:popular   frequency:1
docText:  Classical music is dying
token:classical frequency:1
token:music     frequency:1
token:is        frequency:1
token:dying     frequency:1
```

3. Compute the cosine similarity

In this section, I first filtered the stopwords, then I used the residue words between each question sentence and answer sentence pair to construct several dictionary. And I recalculated the frequency vector. Then I calculated the cosine distance between the two vectors which can indicate the similarity between the question and answer sentence.

After all answer sentences are calculated, I rank them and print them on the console.

```
rel=0   qid=1   Sentence:Pop music has absorbed influences from most other genres of popular music      Score=0.2672612419124
rel=1   qid=1   Sentence:Classical music may never be the most popular music    Score=0.45226701686664544
rel=0   qid=1   Sentence:Everybody knows classical music when they hear it      Score=0.35355339059327373
rel=0   qid=2   Sentence:Old wine and friends improve with age  Score=0.0
rel=0   qid=2   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.0
rel=1   qid=2   Sentence:Climate change and energy use are two sides of the same coin.  Score=0.3061862178478973
rel=1   qid=3   Sentence:The best mirror is an old friend       Score=0.50709255283711
rel=0   qid=3   Sentence:My best friend is the one who brings out the best in me        Score=0.4338609156373123
rel=0   qid=3   Sentence:The best antiques are old friends      Score=0.18257418583505536
rel=0   qid=4   Sentence:Wear a smile and have friends; wear a scowl and have wrinkles  Score=0.22360679774997896
rel=1   qid=4   Sentence:If you see a friend without a smile, give him one of yours     Score=0.17213259316477408
rel=0   qid=4   Sentence:Behind every girls smile is a best friend who put it there     Score=0.28867513459481287
rel=0   qid=5   Sentence:Old wine and friends improve with age  Score=0.11952286093343936
rel=0   qid=5   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.05773502691896257
rel=1   qid=5   Sentence:Old friends are best   Score=0.15811388300841897
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken: 1.186
```

4. Compute the MMR

Then I calculated the MMR of the ranked sentences.

5. Design a better retrieval system

When trying to improve the performance of this system, I made two mainly things here.

5.1 Add Dice and Jarcard Similarity

In the similarity calculating phase, I add Dice and Jarcard similarity to help make it more precisely. The following four pictures are results of using Cosine, Dice, Jarcard and using (Cosine+Dice+Jarcard)/3. We can see the score changes although the ranking didn't change.

```
rel=0   qid=1   Sentence:Pop music has absorbed influences from most other genres of popular music      Score=0.2222222222222222
rel=1   qid=1   Sentence:Classical music may never be the most popular music    Score=0.4
rel=0   qid=1   Sentence:Everybody knows classical music when they hear it      Score=0.3333333333333333
rel=0   qid=2   Sentence:Old wine and friends improve with age  Score=0.0
rel=0   qid=2   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.0
rel=1   qid=2   Sentence:Climate change and energy use are two sides of the same coin.  Score=0.3
rel=1   qid=3   Sentence:The best mirror is an old friend       Score=0.5
rel=0   qid=3   Sentence:My best friend is the one who brings out the best in me        Score=0.36363636363636365
rel=0   qid=3   Sentence:The best antiques are old friends      Score=0.18181818181818182
rel=0   qid=4   Sentence:Wear a smile and have friends; wear a scowl and have wrinkles  Score=0.20689655172413793
rel=1   qid=4   Sentence:If you see a friend without a smile, give him one of yours     Score=0.16666666666666666
rel=0   qid=4   Sentence:Behind every girls smile is a best friend who put it there     Score=0.2857142857142857
rel=0   qid=5   Sentence:Old wine and friends improve with age  Score=0.11764705882352941
rel=0   qid=5   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.05
rel=1   qid=5   Sentence:Old friends are best   Score=0.14285714285714285
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken: 1.24
```

```
rel=0   qid=1   Sentence:Pop music has absorbed influences from most other genres of popular music      Score=0.125
rel=1   qid=1   Sentence:Classical music may never be the most popular music    Score=0.25
rel=0   qid=1   Sentence:Everybody knows classical music when they hear it      Score=0.2
rel=0   qid=2   Sentence:Old wine and friends improve with age  Score=0.0
rel=0   qid=2   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.0
rel=1   qid=2   Sentence:Climate change and energy use are two sides of the same coin.  Score=0.17647058823529413
rel=1   qid=3   Sentence:The best mirror is an old friend       Score=0.3333333333333333
rel=0   qid=3   Sentence:My best friend is the one who brings out the best in me        Score=0.2222222222222222
rel=0   qid=3   Sentence:The best antiques are old friends      Score=0.1
rel=0   qid=4   Sentence:Wear a smile and have friends; wear a scowl and have wrinkles  Score=0.11538461538461539
rel=1   qid=4   Sentence:If you see a friend without a smile, give him one of yours     Score=0.09090909090909091
rel=0   qid=4   Sentence:Behind every girls smile is a best friend who put it there     Score=0.16666666666666666
rel=0   qid=5   Sentence:Old wine and friends improve with age  Score=0.0625
rel=0   qid=5   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.02564102564102564
rel=1   qid=5   Sentence:Old friends are best   Score=0.07692307692307693
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken: 1.252
```

```
rel=0   qid=1   Sentence:Pop music has absorbed influences from most other genres of popular music      Score=0.20482782137821554
rel=1   qid=1   Sentence:Classical music may never be the most popular music    Score=0.3674223389555485
rel=0   qid=1   Sentence:Everybody knows classical music when they hear it      Score=0.29562890797553565
rel=0   qid=2   Sentence:Old wine and friends improve with age  Score=0.0
rel=0   qid=2   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.0
rel=1   qid=2   Sentence:Climate change and energy use are two sides of the same coin.  Score=0.2608856020277305
rel=1   qid=3   Sentence:The best mirror is an old friend       Score=0.4468086287234811
rel=0   qid=3   Sentence:My best friend is the one who brings out the best in me        Score=0.33990650049863275
rel=0   qid=3   Sentence:The best antiques are old friends      Score=0.1547974558844124
rel=0   qid=4   Sentence:Wear a smile and have friends; wear a scowl and have wrinkles  Score=0.18196265495291075
rel=1   qid=4   Sentence:If you see a friend without a smile, give him one of yours     Score=0.14323611691351057
rel=0   qid=4   Sentence:Behind every girls smile is a best friend who put it there     Score=0.2470186956585884
rel=0   qid=5   Sentence:Old wine and friends improve with age  Score=0.09988997325232292
rel=0   qid=5   Sentence:With clothes the new are the best, with friends the old are the best   Score=0.04445868418666274
rel=1   qid=5   Sentence:Old friends are best   Score=0.12596470092954623
 (MRR) Mean Reciprocal Rank ::0.8666666666666668
Total time taken: 1.189
```

5.2Reconstruct the Stopword Filter Process

In this phase, I considered that the Process() method of the uima will run many times for reading each sentence in. So I remove the stopword phase from the DocumentVectorAnnotator class to the collectionProcessComplete() method in RetrievalEvaluator class. Because this will read in the stopword method for just once instead of each time one sentence is readed.

```
Total time taken: 1.167
```

```
Total time taken: 0.871
```

We can see that the performance of the total time have decreased because of this work.