Report

NAME: Qiwen Zhu
ID: qiwenz

One interesting thing I got is that when I tried to create a global dictionary, I first used arraylist to store the words. However, it causes a lot of errors when I try to get its length. Then I found that HashSet is perfect for building a global dictionary since HashSet can automatically judge whether there are duplicates and store only once for each word.

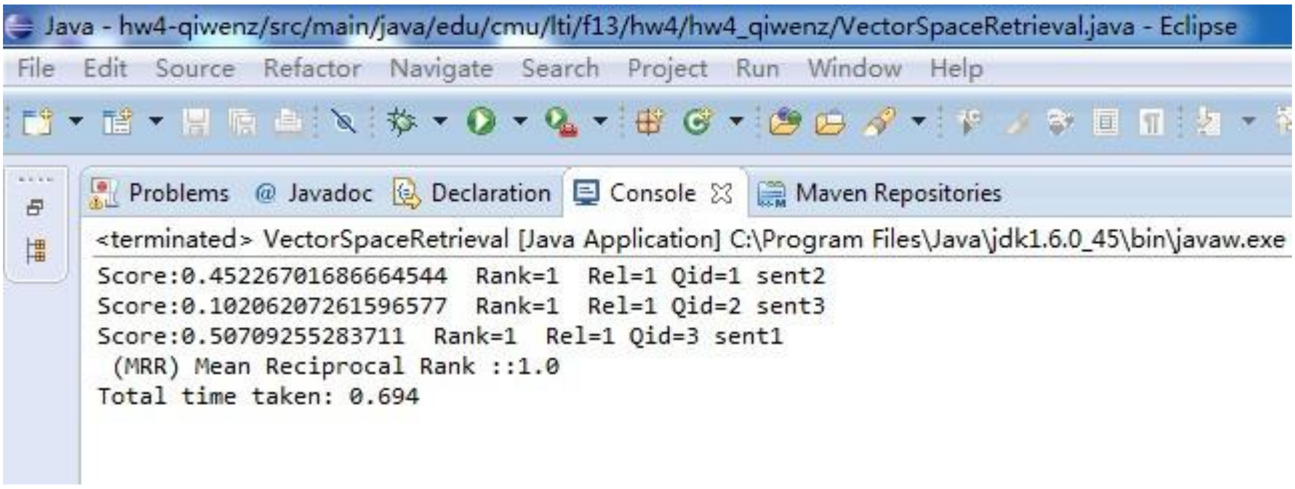
As to whether we need the global dictionary, I think building a dictionary for the two to be compared sentences each time works. However, whether this method can improve the efficiency depends on the size of the global dictionary and the volume of lines.

Finally, I don’t understand why we need to convert the ArrayList to FSList (jcasType need?) in DocumentVectorAnnotator and then convert the FSList to ArrayList immediately in RetrievalEvaluator. What if I modify the typesystem “Document.java”. I wonder if I can change “public void setTokenList(FSList v)” to ” public void setTokenList(ArrayList v)” and what should I be careful about.

Error Analysis

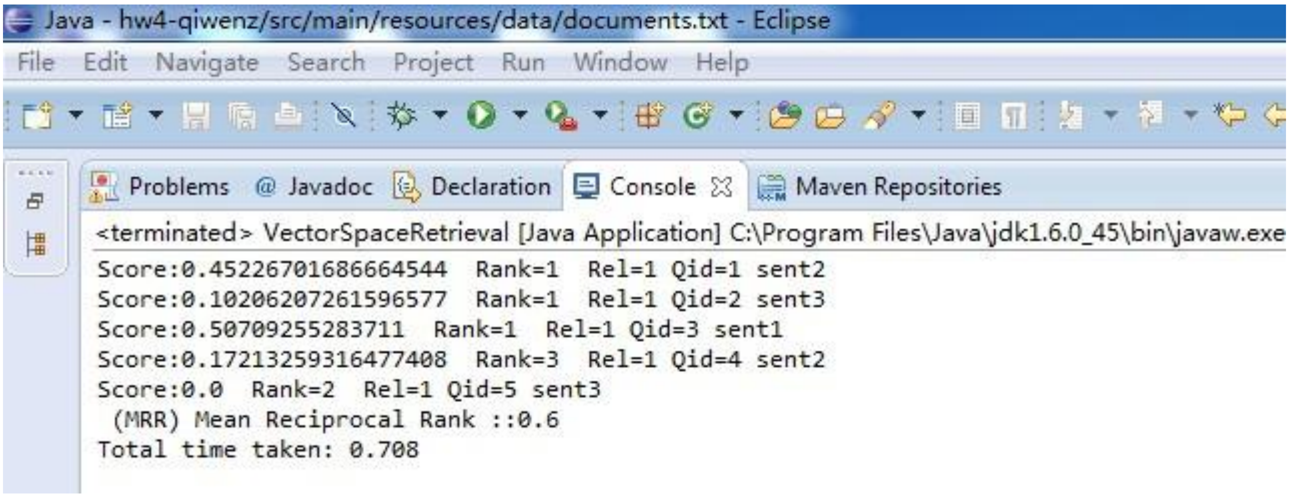
Result:

1.For the first version testing data, I got the result that MRR=1.0 as followed:



```
<terminated> VectorSpaceRetrieval [Java Application] C:\Program Files\Java\jdk1.6.0_45\bin\javaw.exe
Score:0.45226701686664544  Rank=1  Rel=1 Qid=1 sent2
Score:0.10206207261596577  Rank=1  Rel=1 Qid=2 sent3
Score:0.50709255283711  Rank=1  Rel=1 Qid=3 sent1
 (MRR) Mean Reciprocal Rank ::1.0
Total time taken: 0.694
```

2.For the updated version testing data, once I appended the new lines to the documents.txt, the program could not run. The console hints me that there are some errors in the DocumentReader java program which is provided by TAs. Even when I delete all the program I wrote and just left the initial program imported from Maven project, it still could not run. Finally, I copied the original testing data and modified them to the updated testing data and then it run well.


Java - hw4-qiwenz/src/main/resources/data/documents.txt - Eclipse
File Edit Navigate Search Project Run Window Help
Problems @ Javadoc Declaration Console Maven Repositories

```
<terminated> VectorSpaceRetrieval [Java Application] C:\Program Files\Java\jdk1.6.0_45\bin\javaw.exe
Score:0.4522670168664544  Rank=1  Rel=1 Qid=1 sent2
Score:0.10206207261596577  Rank=1  Rel=1 Qid=2 sent3
Score:0.50709255283711  Rank=1  Rel=1 Qid=3 sent1
Score:0.17213259316477408  Rank=3  Rel=1 Qid=4 sent2
Score:0.0  Rank=2  Rel=1 Qid=5 sent3
 (MRR) Mean Reciprocal Rank ::0.6
Total time taken: 0.708
```

Causes of Errors

Cannot recognize different forms (e.g. singular and plural form or capital an small letter) of a word

Let's take question 4 as an example. If the word “friends” in the question sentence can be regarded same as “friend” which appears in answer 2, the score of answer 2 which is the correct answer will be higher. Similar things happened in question 5, such as “old” and “Old”, “friend” and ”friends”.

Cannot recognize word with punctuation

Again, we take question 4 as an example. If the word “smile” in the question sentence can be regarded same as “smile,” which appears in answer 2, the score of answer 2 which is the correct answer will be higher.

Cannot recognize synonyms

Also, we take question 4 as an example. If the word “a” in the question sentence can be regarded same as “one” which appears in answer 2, the score of answer 2 which is the correct answer will be higher.

Cannot recognize reference pro-forms

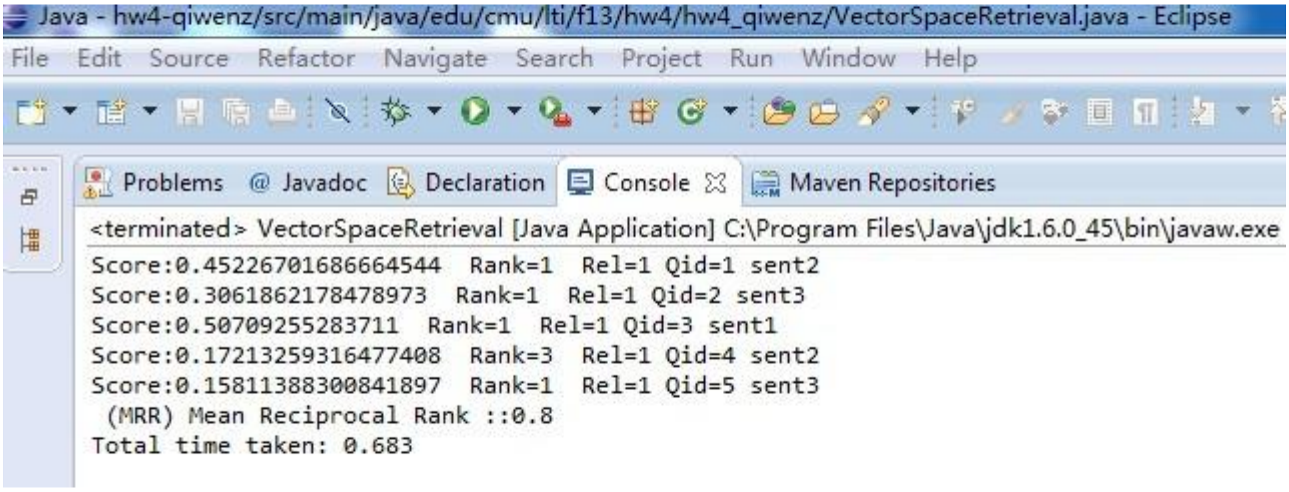
As before, we take question 4 as an example. If the word “him” in the answer 2 sentence can be regarded same as “friend” which is its referred-object, the score of answer 2 which is the correct answer will be higher.

Cannot recognize unmeaningful word

In most time, articles (e.g. “a” “an” “the”) do not affect the meaning of a sentence. If we can delete these kind of word, the result would be more reasonable.

Improvement

It is easy to convert all the capital letter to small letter in java (using .toLowerCase()). And this help to get better rank for question 5.



Java - hw4-qiwenz/src/main/java/edu/cmu/lti/f13/hw4/hw4_qiwenz/VectorSpaceRetrieval.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Problems @ Javadoc Declaration Console Maven Repositories

```
<terminated> VectorSpaceRetrieval [Java Application] C:\Program Files\Java\jdk1.6.0_45\bin\javaw.exe
Score:0.45226701686664544  Rank=1  Rel=1 Qid=1 sent2
Score:0.3061862178478973  Rank=1  Rel=1 Qid=2 sent3
Score:0.50709255283711  Rank=1  Rel=1 Qid=3 sent1
Score:0.17213259316477408  Rank=3  Rel=1 Qid=4 sent2
Score:0.15811388300841897  Rank=1  Rel=1 Qid=5 sent3
 (MRR) Mean Reciprocal Rank ::0.8
Total time taken: 0.683
```

Other Thoughts of How to Improve

Dealing with synonyms and different forms of a word

Building a dictionary (or reference) for nouns and verbs. When we find a noun or a verb, we will search the dictionary and try its other forms or synonyms. Different forms of a word of synonyms can be regarded as the same when we do the comparison.

Dealing with punctuation

Using other method to split the sentence and get token. For example we use regular expression such as “[a-zA-Z]+” to get token, then the punctuation will not be contained within a token. Another advantage that ignore the punctuation is that expression such as “XXX’s” now can be regarded as two token “XXX” and ”s”, and “XXX” may well appears before.

Dealing with unmeaningful word

Building a dictionary of unmeaningful word. Before we put a word in token list, we tranverse the dictionary to check whether it is an unmeaningful word. We discard this word if it appears in the dictionary.

Dealing with reference pro-forms

From my point of view, it is hard to achieve this. Maybe machine learning can help to solve this problem.