11791 Design and Engineering of Intelligent Information Systems

Fall 2013 Assignment 4

Kuo Liu
kuol@andrew.cmu.edu

October 22, 2013

Task 1

After dropping stop-words and changing all the words to the lowercase, I get the following result. The output for task1 is as follows:

```
Score:   0.6123724356957945   rank=1   rel=1   qid=1   Classical music may never be
the most popular music
Score:   0.4629100498862757   rank=1   rel=1   qid=2   Climate change and energy
use are two sides of the same coin.
Score:   0.5  rank=2   rel=1   qid=3   The best mirror is an old friend
(MRR) Mean Reciprocal Rank ::0.8333333333333334
```

Task2

Error Analysis: Among the three test cases, one fails to rank to the relevant document to the first place. The reason is that after deleting the stop-words, the term vectors for query and the corresponding texts of case3 are:

```
Query: {oneself=1, one's=1, best=1, friend=1}
Text1: {mirror=1, old=1, best=1, friend=1}          relevant
Text2: {one=1, best=2, brings=1, friend=1}          irrelevant but ranked first
Text3: {old=1, friends=1, best=1, antiques=1}
```

We can figure out that that the term frequency for best is 2 in Text2, while that is 1 in Text1. This point leads to the bad result. But when we turn to see the original texts, we can figure out that old friend is the synonym of best friend (i.e. we can deem old and best as the same).
We can figure out the reason for this fail here which is we fail to consider synonyms.
So, we can add one part to do the analysis of synonyms. But since synonyms dictionaries or tools such as wordnet are not available this time, I just add one rule for simplicity. If we come up with “old friend”, we deem the word “old” as “best”. Here, we can just consider “best” as the candidate of one group of synonyms.

Actually, I have considered several ways to do this kind of job.

1. Use Statistical Models like BM25 or Language Models to do this kind of job. But considering that the documents are two short which results in too little statistical properties to use. I didn't use it.
2. Use several ways to compute different similarities and sum them up, but it comes out that this way is also useless to improve MRR.
3. Use different ways to compute similarities and each way should vote for the text which gets the highest corresponding score. After the vote, we do a simple bagging. This way is also useless to improve MRR.

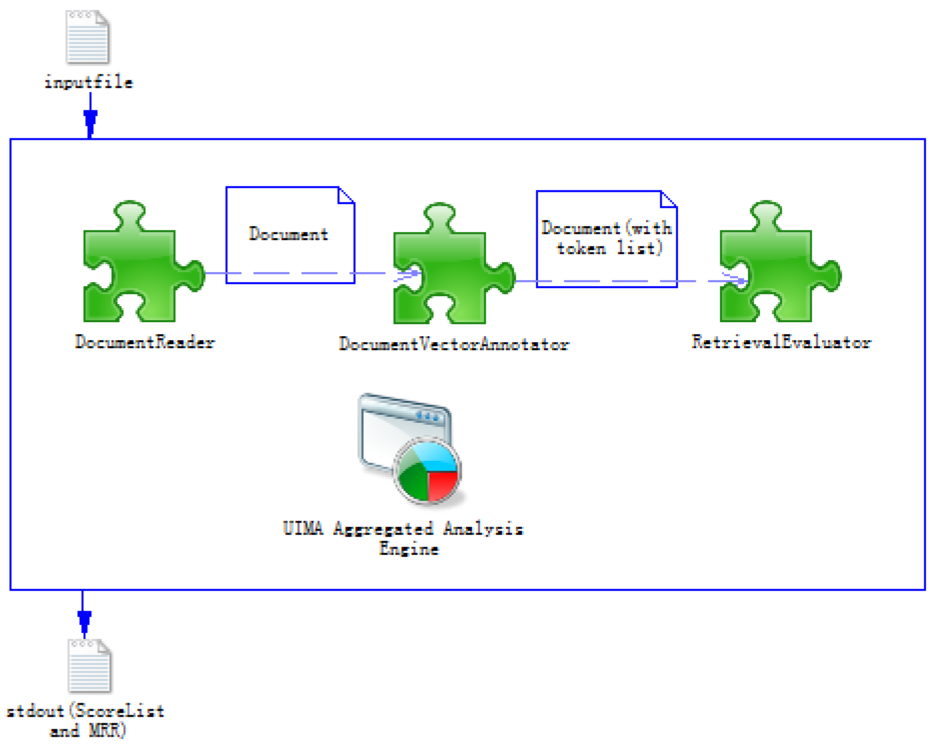
At last, I find out that the fundamental problem is about semantic analysis, in our case, it is synonym analysis.

Design and Implementation

The design pattern is just the same as the given template and I implemented the corresponding part and did some necessary changes.

There are two types, Document and Token, in my homework4 project.

The plot for elements and dataflow is as follows:


inputfile
Document
Document(with token list)
DocumentReader
DocumentVectorAnnotator
RetrievalEvaluator
UIMA Aggregated Analysis Engine
stdout(ScoreList and MRR)

Bonus

I tried cosine similarity, dice coefficient and Jaccard coefficient on the test dataset using the given stop-word list and my system. The result is listed as follows.

Similarity Computation	**Cosine Similarity**	**Dice Coefficient**	**Jaccard Coefficient**
MRR	1.00	1.00	1.00

The result of the experiment can’t display the relative advantages or disadvantages among the three methods of similarity computation since the dataset is too small. Here comes the analysis for each method.

For Dice Coefficient and Jaccard Coefficient, the two methods are very similar since they both consider the degree of overlap as the measurement of similarity. The disadvantage is that they both take each element as equally important which can be a problem. Since different elements should have different weights, assigning the same weight to each element is not suitable in the usual case. Cosine Similarity corrects this disadvantage by considering the weight for each element. But it also has some disadvantages. It just considers the angle between two vectors. In this case, two vectors with the same scale are considered to be the same which can lead to problems sometimes. For example, can we deem the two vectors <1, 2> and <2, 4> as the same? In some cases, the answer may be not.