11-791 Design & Engineering of Intelligent Information Systems

11-791 Design & Engineering of Intelligent Information Systems

Hw3-Report

AndrewID: Chongshm

For The Task 1

1. The Design Aspects

1.1 UML Design Diagram

The Design Pattern, which I used as the following UML diagram, I revise two of the classes in the original pattern, and add new data type class called the "storage" to store all the useful information. The Document reader is responsible for extracting the qid , rel, etc. information, the annotator is in charge of extracting and splitting the words in the sentence and process these sentences .
The job of the casconsumer is to calculate the cosine similarity and Mean reciprocal rank. The detailed explanations are in the next section.

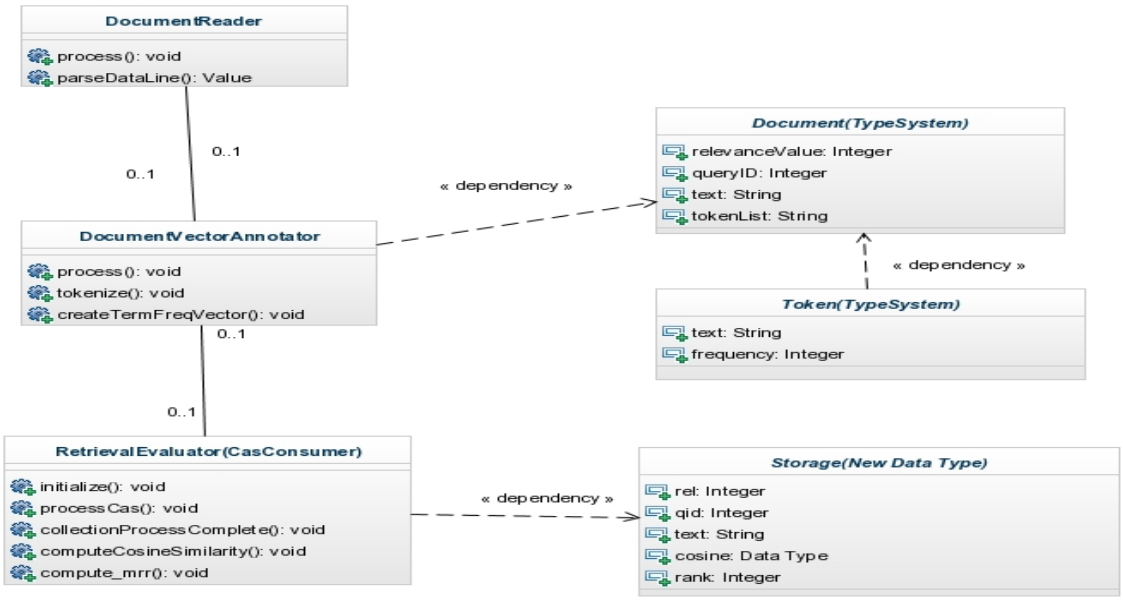

Figure 1. The UML diagram of Task 1

1.2 Design of Document Vector Annotator:

I did not change the structure of this Annotator. And I just add a function of this annotator, which is used for extracting each word in a sentence. Moreover, I followed the instruction in the assignment PDF with using the dumb tokenize to

complete this function.

1.3 New Data Type Design:

I used two different type systems in this homework. The type system, which is called the Annotation, I create it with 4 features as following:

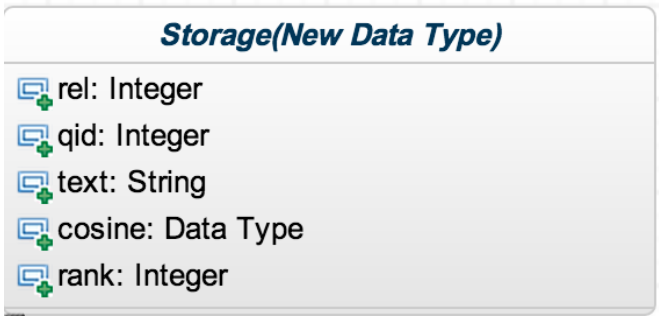

Figure 2. Data type for storing the information of each sentence.

(1)rel, which stores the a binary relevance assessment for a specific sentence of document.
(2)qiu, which represents the a query id.
(3)text, which record the content of a specific sentence .
(4)cosine, which is responsible for storing the cosine similarity of each sentence.
(5)rank, which is record the sentence' s rank based on cosine similarity.

1.4 CasConsumer Design

The CasConsumer in this task is the most significant part. In the CasConsumer, I have placed twos functions for computing cosine similarity and Mean Reciprocal Rank(MRR) respectively.

1.4.1 Cosine Similarity

For computing the cosine similarity, I assigned this function during processing the Cas, When there existed a new cas (sentence), I called the tokenlist function, and using two hashmaps to store the frequency and content of each word. Two hashmaps were designed for query and the relevant sentence below it respectively. When the Cas moved to the next query, this function would clear these two hashmaps and begin to record the current query and its sentence. The key of these two hashmaps was the word, the value is it frequency. When this function complete to calculate the cosine similarity for each sentence, these information, including the query id, binary relevance assessment, cosine similarity, and text content will be

recorded in an Storage type arraylist. According to each query, there existed another hashmap, “questionmap”, to store the same query id’s arraylist. Therefore, the key would be the query id, the value should be the arraylist.

After finishing processing all the Cases, the rank function will start to do its job. By going through all the keys in the questionmap respectively, we could get the rank of all sentences in a query. The rank was calculated by comparing the cosine similarity of each sentence in a query. Meanwhile, the rank will be saved in its arraylist. Then, the result of sentence whose relevance assessment is 1 transferred to store in another “storage” type arraylist named “output”.

1.4.2 Mean Reciprocal Rank

As the last section the new Data Type “Storage” presented, I used the storage as arraylist to store all the information to compute the Mean Reciprocal rank. With finishing processing all the sentences and utilizing the rank from output arraylist, The Mean reciprocal rank could be calculated by following the equation in the homework3 introduction.

1.5 Final report for the task 1

The MRR result was 0.4735 as following figure.

```
cosine=0.0788 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.2828 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.1508 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.2265 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.1268 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.4375
Total time taken: 1.273
```

Figure 3. The final result for the task 1.

For The Task 2

1. Major Error Summary and Comparison

1.1 Major Error Types

The error analysis should be divided into 3 major types as Following.

(1). Vocabulary Mismatch

1). The tense difference, such as “purchase” &”purchased”(qid=3), the “scored” & “score”(qid=4), etc. Thus, this kind of tense difference will lead to lower the cosine similarity score of the real relevant sentence.
Targeted Qid: 3, 4, 8, 11, 15, 16
2). The singular & plural difference, such as “point”& “points” (qid=4), etc. Thus, this kind of difference could make the cosine similarity become lower than normal.
Targeted Qid: 1, 4
3). The abbreviation difference, such as “McDonald's”& “McDonald” (qid=18), etc. Thus, this kind of difference could disturb the calculation of cosine similarity.
Targeted Qid: 11, 18
4). The punctuation difference, because of splitting the word by using a space, thus, the punctuation difference will significantly affect the precision of similarity.
Targeted Qid: 1, 2, 6, 12, 15, 16, 17
5). The Capital letter and Lower letter difference, it is possible that the core none was included in the query or sentence as first word, but the cosine similarity function cannot match the same Capital letter and Lower letter.
Targeted Qid: 2

(2). Irrelevant information

1). Preposition & indefinite article distraction, such as the “of”, “the”, etc. Just like the stopword document showed. With counting these meaningless information, the cosine similarity could be raised. However, this useless information will disturb the precision of ranking.
Targeted Qid: 6, 7, 13, 14, 15, 17, 18, 20
2). Same word but different main idea, there exists a situation that the sentence is able to include some of the core words in the query, but the main idea of a sentence does not directly answer the question of query.
Targeted Qid: 1, 2, 3, 4, 5, 6, 7, 8, 11, 16, 17

(3). Information redundancy

Because of too much detail in the sentence, even if the main idea of the sentence is related to the query, the cosine similarity score of the sentence will not be ideal.

Targeted Qid: 1, 2, 4, 5, 7, 8, 12, 13, 14, 15, 17, 18, 19, 20

2. Major Error Improvement

2.1 For Error 1 (Vocabulary Mismatch)

2.1.1 Stemming Algorithm:

According to the introduction of homework3, "untils" package provides the StanfordLemmatizer.java, which could help to process the tense difference and the singular & plural difference. With the Stemtext function, the cosine similarity is much higher than previous.

Improvement: 0.4375 -> 0.5500

```
cosine=0.0727 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.4243 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.3015 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.2265 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.2417 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.5500
Total time taken: 1.002
```

Figure 4. The result of using StanfordLemmatizer alone.

2.1.2 Tokenization Algorithms:

Meanwhile, the tokenization algorithms also should be used in dealing with the vocabulary mismatch, especially abbreviation difference and the punctuation difference. Owing to utilize this kind of function the result was improved slightly.

Improvement: 0.4375 -> 0.4958

```
<terminated> VectorSpaceRetrieval [Java Application] /Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Home/bin/java (Oct 21, 2014, 7:08:26 PM)
cosine=0.0788 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.2828 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.2085 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.2621 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.1670 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.4104 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.4958
Total time taken: 1.058
```

Figure 5. The result of using Tokenizer alone.

2.2 For Error 2 (Irrelevant information)

2.2.1 Preposition & Indefinite Article Distraction:

With taking the stopwords.txt as reference, removing all the words in these sentences and queries. It could be seem as a feasible way to overcome the Preposition & Indefinite Article Distraction problem, because “if” “there”, and so on, such these word will not participate in calculation of cosine similarity. Therefore, the performance also achieved obvious improvement.

Improvement: 0.4375 -> 0.4750

```
cosine=0.1179 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.3780 rank=2 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0000 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.0000 rank=3 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.0745 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.1826 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.4750
Total time taken: 1.503
```

Figure 7. The result of removing Stopwords alone.

2.3 Combination of the three methods as above

In the annotator, in order to combine of the three algorithms, I replaced all the punctuations in dealing with the document string. Then, I made the string process with Stemming Algorithm. In the last step, I created a new arraylist to store all the stopwords. Moreover, the processed string should be determine whether it have the same word in the stopword list, if the stopword contains the word, then remove it. If not, save the word into the map. The improvement is significant.

Improvement: 0.4375->0.6625

```
cosine=0.1455 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.4243 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.4170 rank=2 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.2621 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.2789 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.4104 rank=1 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.6625
Total time taken: 1.129
```

Figure 8. The result of combination of three methods.

2.4 Implement method to compare with the cosine similarity

Due to the error 3, we need to figure out a new way to deal with the information redundancy. Thus, not only compare the similarity, but also evaluate the text position. Thus, there exist 2 methods to help us to solve this problem.

2.4.1 Jaccard coefficient method

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

By using this method, we may not get a better Mean Reciprocal Rank (MRR) than the cosine similarity. It just simply gets the intersection set and the union set.

Implement: 0.4375->0.4417

```
cosine=0.3333 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.0385 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature r
cosine=0.1538 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0690 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing areas
cosine=0.0278 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that
cosine=0.0556 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurs
cosine=0.1429 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living
(MRR) Mean Reciprocal Rank ::0.4417
Total time taken: 1.038
```

Figure 9. The result of using Jaccard coefficient method.

2.4.2 Tversky coefficient method

The Tversky index, named after Amos Tversky, is an asymmetric similarity measure on sets that compares a variant to a prototype. The Tversky index can be seen as a generalization of Dice's coefficient and Tanimoto coefficient. For sets X and Y the Tversky index is a number between 0 and 1 given by

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|},$$

By using this method, we could get a better Mean Reciprocal Rank (MRR) than the cosine similarity. By considering the similar text position, the result should be more

persuasive than simple comparison of the text similarity. Therefore, this method raised the MMR index significantly.

Implement: 0.4375->0.5333

```
cosine=0.4000 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.0400 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature
cosine=0.2000 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0294 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing are
cosine=0.0526 rank=3 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach th
cosine=0.0588 rank=2 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehu
cosine=0.0968 rank=3 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is livi
(MRR) Mean Reciprocal Rank ::0.5333
Total time taken: 0.962
```

Figure 10. The result of using Tversky coefficient method.

2.4.3 Dice coefficient method

The Sørensen–Dice index, also known by other names (see Names, below), is a statistic used for comparing the similarity of two samples.

$$QS = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A| + |B|}$$

At last, I also realized the dice-coefficient method. The result of this method is the same as the Jaccard-coefficient algorithm.

Implement: 0.4375->0.4417

```
cosine=0.3333 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.0385 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature r
cosine=0.1538 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.0690 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal   temperatures in major growing areas
cosine=0.0278 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that
cosine=0.0556 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurs
cosine=0.1429 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living
(MRR) Mean Reciprocal Rank ::0.4417
Total time taken: 1.038
```

Figure 11. The result of using Dice coefficient method.

3.Algorithm Design

From the result of the task 1, the single cosine similarity calculation could not return the reasonable rank for the document. Thus, in the task 2, I summarized all the possible error types in analyzing the sentences of document.txt. First, I extracted all the qids which are relevant to the error 1 and 2. By using the method of stemming algorithms and tokenization algorithms, the rank of qids, which had error 1, was

11-791 Design & Engineering of Intelligent Information Systems

improved significantly. After that, for the error 3, I utilized another methods called "Jaccard coefficient" to solve. Focusing on the replacement the cosine similarity ranking way, the Tversky coefficient reveal its ability.

4. Advantages and Disadvantages

The advantage of the cosine similarity with tokenization algorithm and stemming algorithm is obvious. The value of MRR increases significantly. However, utilizing the Jaccard coefficient and Tversky coefficient method could not reach such high level. Thus, it means the major error type is not the mismatching position. But, compare to the task 1 the Tversky coefficient also improved the performance a lot. Therefore, it is useful to utilize Tversky coefficient. Meanwhile, the disadvantage of the improvement is that this function just implemented several general methods, and did not combine these methods. Distributing the weight to each method could be a feasible way.

5. Summary

In this task, the most difficult part for me was to find the rank method to get the most reliable results and process the raw data as precise as possible. Comparing and evaluating these information retrieval methods is the valuable thing I learned from this project. With adding different methods to ranking the sentence, the precision will improve simultaneously. I hope the IR experience could be utilized in the coming team homework.