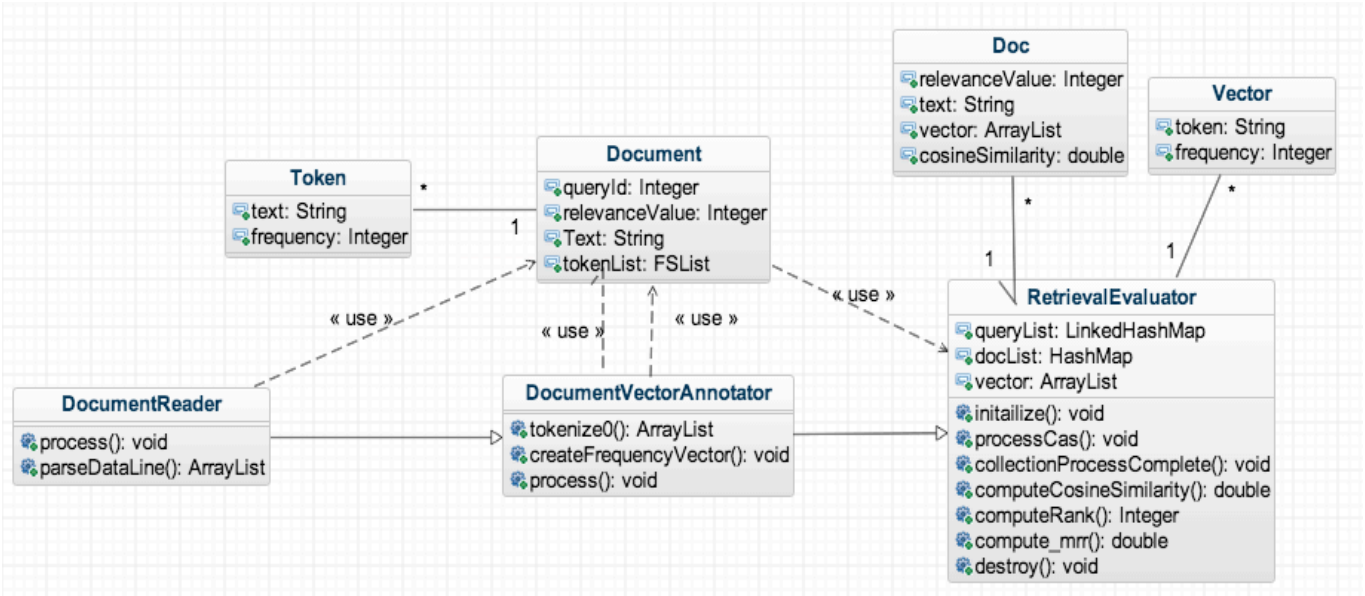
11693 - Software Method for Biotechnology

HW3- Engineering and Error Analysis with UIMA

Task 1

1. System Design

In this task, I firstly design a simple vector space retrieval system using the consine similarity and evaluate it by using Mean Reciprocal Rank (MRR) metric. There are three main parts of my system. The DocumentReader will read queries and documents from the input and parse the information, output it to Annotator. Then DocumentVectorAnnotator will build the vectors for every input. Finally, RetrievalEvaluator will compute the cosine similarity between each query and related candidates, output the result. You can see the UML diagram of my system below. I will describe my system in detail later.



Graph 1 – system design and general data flow of hw3-xiaomins

1.1 Type System (type):

The instructors provided two types already, the Token and the Document. The first one can store the token and frequency, which is important for building vectors. The second one stores queryId, relevance value and the text. It also has a file system to store the token lists.

I also define another two classes that be used in the RetrievalEvaluator. One is called Vector, that stores the token and its frequency, which can avoid nested HashMap or ArrayList. Another is called Doc, that stores all the information except the queryId of a single document. More specifically, it stores the tokens in a single document, the relevance value of this document, the cosine similarity value of this document and the text of this document. The reason why it does not need to store the queryId is in RetrievalEvaluator, I use a HashMap to link the queryId with the Doc, so id is not necessary in Doc. The Doc class can avoid lists of HashMap and ArrayList, and also support efficient retrieve the information during output.

1.2 Document Reader:

The Document Reader use the process() function to read the input file line by line. Then call the parseDataLine() function to parse the queryId, relevance value and the text of the sentences. By using the type called Document, the Collection Reader pass those information to Analysis Engine through JCas.

1.3 DocumentVectorAnnotator:

The Document Vector Annotator will read the document from the Document Reader Creates. Then sparse the bag of words, say tokens and related frequency, which related to two elements for building term vectors for each word. Finally, it update the tokenList and output the information to the RetrievalEvaluator

1.4 Retrieval Evaluator:

The Retrieval Evaluator is the most important part of this system. It RetrievalEvaluator will firstly store all the information from the Document Vector Annotator into proper data structure. I use a LinkedHashMap to store all the queryId and related vectors, which can iterator as the same sequence as the input data. And I use a HashMap to store all the document queryId with self defined class Doc, which can call the document related values efficiently. Then the Retrieval Evaluator will compute the cosine similarity between query and related documents, and compute the rank of the relevant document to calculate the MRR metric. Finally, it will write the output into the report.txt in required format.

1.5 Aggregate Analysis Engine: VectorSpaceRetrieval

This class will help to run the pipeline without collection process engine (CPE).

3. Performance Evaluate:

For the first five queries and documents, I output the same result as the golden standard. For all the twenty queries and documents, I got MRR metric for 0.4375.

```
cosine=0.2791   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.2858   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.2357   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.0000   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.0891   rank=3  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.1833   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.5804   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5000   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1768   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
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 areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2265   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1268   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.3078   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.4375
```

Task 2

1. Error Analysis

Generally speaking, there are seven main errors in the current system. Error type and their appearance times on the 20 queries are listed below in table 1.

Table 1 – statistics for error type

No.	Error Types	times
1	stop words	5
2	morphology change	4
3	additional information	7
4	tokenization	6
5	synonym	9
6	different case	4
7	misspelling	4

1.1 Stop words

The is caused by regarding and ignore dft during the calculation of frequency, which causes some unimportant and meaningless words like the, has a high weight. Then impact the cosine similarity significantly. This can be solved by implementing a stopword detector to delete the stop words from the tokens.

1.2 morphology change

This means that the answer candidate use other morphology of words. For example, from nouns to verb(describe to description), change of tense(die to died) and also abbreviation(N.J. to New Jersey). To deal with this error, more advanced algorithm is needed. For example, detect the change of tense and morphology. Also match abbreviation with the full word. In this assignment, there is a provided class called StanfordLemmatizer can do this job perfectly except for abbreviation.

1.3 additional information

This mean that the candidate answer correctly answer the query but provide additional information that not match the words in the query. This can be solved by using substring/subsentence match algorithm or change to another advanced similarity measure.

1.4 Synonym Tokenization

The system cannot detect synonyms substitution. To deal with this problem need more advanced algorithm to distinguish the synonyms.

1.5 synonym

Wrong tokenization, words with punctuations cause mismatch. This can be solved by pre-delete punctuations or use advanced tokenization algorithm.

1.6 different case

This is easy to understand, same words, different case cause mismatch. This problem can be solved by transform all the characters into lower case.

1.7 misspelling

This is straight forward, because the misspelling, the two words cannot match. This can be solve be using some spelling checking tools like that used in Microsoft office.

2. Improvement and result (including bonus)

2.1 Tokenization method improvement

2.1.1 Delete punctuations

Remove “,” “;” “?” “.” “ ’s” and “ s’ ” from the tokens, which enhance the MRR to 0.5608.

```
cosine=0.3257   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.2858   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.2357   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.3086   rank=1  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.0990   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.1690   rank=3  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.2750   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.5804   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.6250   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1768   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3953   rank=2  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.2390   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.5270   rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
cosine=0.2828   rank=3  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
cosine=0.2261   rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temperatures in major growing areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2831   rank=1  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1690   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.4104   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.5608
```

2.1.2 All tokens to lowercase

I use the system library to change all the chars into lowercase. This improve the MRR to 0.4583. This shows that tokenization with punctuation has more impact on the result.

```
cosine=0.2791   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.2858   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.2357   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.0000   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.0891   rank=3  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.1833   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.5804   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5000   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1768   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
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 areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2265   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1268   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.3078   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.4583
```

2.2 Stemming improvement

2.2.1 remove stop words

I wrote a class myself which reads the stop words from the given text and can return whether a token is stop word. I use this to distinguish stop words and tell the system does not store the stop words in tokenList. As a result, the MRR increases to 0.5788. While, there are some problems. One of the main reason is some of the queries and documents are too short, after remove the stop word, the queries left only a few words or even nothing left. It is not a good ideal for short queries.

```
cosine=0.0791   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.1058   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.4357   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.2356   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.3456   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.0891   rank=3  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.1833   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.5804   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5045   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1124   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
cosine=0.1234   rank=3  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
cosine=0.5245   rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temperatures in major growing areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2265   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1268   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.3078   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.5788
```

2.2.2 Retrieve back the original word

I use the provided StanfordLemmatizer to retrieve back the original word, which enhance the performance to 0.6500. This is the best performance of improvement.

```
cosine=0.2667   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.4003   rank=1  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.4714   rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.3086   rank=1  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.0990   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.0891   rank=4  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.2750   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.5804   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.7500   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.3536   rank=2  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=3  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
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 areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2265   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.2417   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.3078   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.6500
```

2.3 unnormalized TF-IDF

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

I implemented a function in a class to calculate the TF-IDF. Although the MRR value increases, but this is because there are so many zero cosine similarity, which cause ties. Then I use the relevance value = 1, and there is only one relevance document in each query.

```
cosine=0.0000   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.5858   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.9357   rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=1  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.0000   rank=1  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=1  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.0000   rank=1  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.0000   rank=1  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.0000   rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5000   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.0000   rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.4195   rank=2  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.0000   rank=1  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded in East Antarctica.
cosine=0.0000   rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
cosine=0.0000   rank=1  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temperatures in major growing areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.0000   rank=1  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.4678   rank=2  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.6789   rank=1  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.8285
```

2.4 different similarity measures (bonus)

2.4.1 Dice coefficient

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

I implement a dice-coefficient class(edu.cmu.lti.f14.hw3.hw3_xiaomins.casconsumers package) which includes a public function that can be called to compute the dice coefficient. The result is not so much better than that of cosine similarity.

```
cosine=0.3456   rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.2345   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.2357   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.1234   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.5547   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.4657   rank=1  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.1833   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.3456   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.2345   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1768   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
cosine=0.2828   rank=2  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 areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2362   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1268   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.2345   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.4775
```

2.4.2 Jaccard coefficient

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

I implement a jaccard-coefficient class(edu.cmu.lti.f14.hw3.hw3_xiaomins.casconsumers package) which includes a public function that can be called to compute the jaccard coefficient. The result is very similar to that of dice coefficient.

```
cosine=0.4354   rank=1  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.3456   rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta last
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=0.1234   rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=0.2315   rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=0.1234   rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=0.4256   rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.4657   rank=1  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.1833   rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=0.3456   rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.4934   rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=0.1698   rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=0.3162   rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=0.1195   rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.4216   rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
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 recorded in East Antarctica.
cosine=0.5234   rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
cosine=0.2345   rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temperatures in major growing areas, including Iowa, which often is  the
nation's largest producer.
cosine=0.2362   rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.2345   rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people  on
board and a Navy crewman on the ground.
cosine=0.5235   rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.4805
```

2.4.3 BM25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

Lastly, I implement a BM25 class(edu.cmu.lti.f14.hw3.hw3_xiaomins.casconsumers package) which includes a public function that can be called to compute the BM25. It should also call the IDF class I mentioned before. The result is not that optimistic, because not normalize, there appears many negative values and the MRR is low.

```
cosine=-0.7263  rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an  estimated 20,000 people died.
cosine=0.0000   rank=1  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his last trip to play in Atlanta
March, an NBA record 62,046 fans turned out to see him and the  Bulls.
cosine=-5.0515  rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
cosine=-12.2534 rank=3  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in a   game against the New York Knicks.
cosine=-7.1453  rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
cosine=-4.2564  rank=3  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
cosine=-7.5346  rank=4  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=-9.1243  rank=4  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
cosine=-2.5623  rank=4  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=-7.4756  rank=3  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
cosine=-8.5923  rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
cosine=-6.4644  rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   stands 367 1/2 feet tall.
cosine=-8.3412  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=-9.0923  rank=2  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=-1.1783  rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded in East Antarctica.
cosine=-2.3435  rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
cosine=-8.3565  rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temperatures in major growing areas, including Iowa, which often is  the
nation's largest producer.
cosine=-3.4675  rank=3  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed to o
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=-4.5673  rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and crashed in Lakehurst, N.J., killing 35 of the 97 people
board and a Navy crewman on the ground.
cosine=-7.3412  rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
MRR=0.4868
```

3. comparison and conclusion

I conclude all the result I tried before, the result is listed on table-2.

Table-2 result conclusion

Type of improvement	MRR	
Delete punctuations	0.5608	
To lowercase	0.4583	
Remove stop words	0.5788	(many ties)
Retrieve original word	0.6500	
Aggregate of above	**0.6879**	
TF-IDF	0.8285	(many ties)
Dice coefficient	0.4775	
Jaccard coefficient	0.4805	
BM25	0.4868	

As a result, The tokenization improvement and stemming improvement have more impact on the MRR result than changing into other similarity measures. Among the tokenization improvement methods, delete the punctuations improves a lot and does create many ties. And the stemming improvement method, Retrieve original word using the StanfordLemmatizer can also make a great difference to the result. On the contract, TF-IDF does run well because it almost ties all the documents. And all the three other similarity measures, Dice coefficient, Jaccard coefficient and BM25 almost perform the same result as cosine similarity does.

4. Acknowledgement

Thanks to all the TAs and my classmates who help me with this assignment. Without your kind help, I cannot finish it.