Final Report for Homework 3

Andrew ID: leixiao

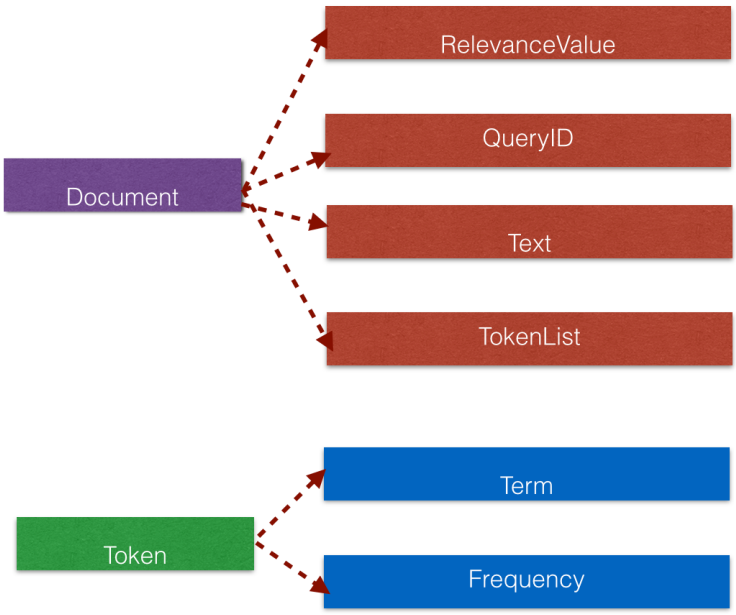
Name: Lei Xiao

Task 1

1. Architecture Design for task 1

(1)Type System

In this part, I designed two types in it, which include Document and Token.

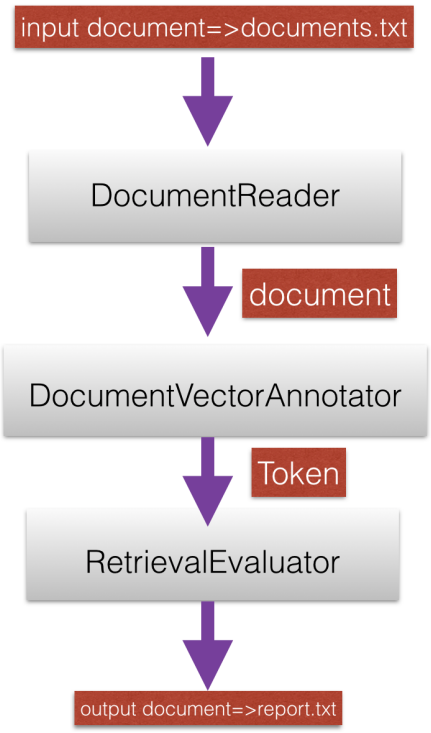


Document has four functions, namely RelevanceValue, QueryID, Text and TokenList. This type is used to deal with the mission that when reading the input file, the CollectionReader can automatically turned the input file into several sentences.

Token has two functions, namely Term and Frequency. This type is used in the annotator. Token is the analyzing result of the Document type, and turns the token list into the tokens, which is separated from the space, and then sends it to the casConsumer, which could put the content of CAS in the output document, and frequency is the times that the term exists in the sentence.

(2)Engineering Process

In this part, I will introduce the whole architecture of the process.



The graph shows the process of dealing with data and specific types in each component of the system.

(3)Design Pattern

As can be seen from the picture, the whole design of the system can be divided into three parts, namely DocumentReader, DocumentVectorAnnotator and RetriecvalEvaluator. Each component has its own functions and roles in this system.

1) DocumentReader,whose function is to identify the query id and relevant scores, and then give the value to RelevanceValue, QueryID, Text and TokenList. The main idea of this part is to get the initial format (Sentence) from the input file and transfer it to the annotator to process it later.

2) DocumentVectorAnnotator, which is to identify the word and word frequency in each sentence and store it in the token.type. And using a basic white-space tokenizer, which deliberately does not split on punctuation.

3)RetriecvalEvaluator, whose function is to compute the similarity and performance metric of the collection. And this part is the most important part in this task 1.

(1)About calculating the cosine similarity, we use the formula as follows:

$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

(2)About calculating the MRR:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i},$$

(3) To get the rank of each sentence, we need to calculate the cosine similarity of each document and its query, and then we could rank them.

Task 2

1.Error Analysis

The original result from the task 1 is as follows, whose tokenization algorithms is only using simple white-space tokenizer.

Result for only using simple white-space tokenizer

```
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=0.4374999999999999
```

As can be seen from the graph above, we may know that only one most relevant answer is selected (which sentence is tagged yellow) from 20 queries, so the MRR is very low.

To show the errors clearly, we could classify the errors into different classes, and improve them separately.

	error type	number of queries that related to this error	specific queries	explanation	example
1	vocabulary mismatch (different tense)	4	qid=2,3,4,16	This error is caused by the different tense.Using the StanfordLemmatizer, we could easily solve this problem.	The error is just like not to determine the different from doing, did, done and do.
2	punctuation mark	3	qid=11,14,20	This error is caused by the key word which is followed by punctation.	The error is just like A, could be seen as a token, however A is actually the real key word.
3	lowercase and uppercase	4	qid=2,3,4,16	This error is caused by recognizing the difference between lowercase and uppercase.	For example, in this sentence,”Music a good music.”. In this case, Actually, Music is the same as music, however, with the basic method, it tells that they are the different.

	error type	number of queries that related to this error	specific queries	explanation	example
4	error from unnecessary words(stop words, such like is, am, are, have)	2	qid=5,7	This error is caused by over consideration about the stop words(such like is, am,are,the), which has no meaning function in the sentence.	For example, in this sentence,"The car has the red the girl like." The word 'the' may contain a big ratio in calculating the cosine similarity, which will cause mistakes.
5	vocabulary order			This error is caused by semantic analysis, which could not be understood by just tokenizing the words.	For example, in this sentence,"A is B", however when calculating its cosine similarity with "B is A", it will get the full score. Although they have the same words, they have quite the different meaning.

2. Improving performance

(1) Better stemming algorithm

This error is caused by the different tense.Using the StanfordLemmatizer, we could easily solve this problem. The error is just like not to determine the different from doing, did, done and do, when we ignore them, we could get a better performance.

- Using the provided wrapper for the Stanford lemmatizer (package util.StanfordLemmatizer) could solve the different-tense related problems and ignore the different forms of each word.

And the result is as follows:

Result for using StanfordLemmatizer

```
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=0.55
```

The sentences tagged green is optimized by this Stanford lemmatizer.

Also, the following is using the qid=3 as an example to fully illustrate this stemming algorithm.

When it could ignore the difference between “purchase” and “purchased”, the system could have better performance.

Original rank for qid=3

```
cosine=0.0000 rank=0 qid=3 rel=99 In which year did a purchase of Alaska happen?
cosine=0.4714 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.2673 rank=2 qid=3 rel=0 1867 - U.S. President Andrew Jackson proclaims treaty for purchase of Alaska from Russia.
cosine=0.4216 rank=1 qid=3 rel=0 William Seward negotiated a purchase of Alaska for $7.2 million.
```

updated rank for qid=3

```
cosine=0.0000 rank=0 qid=3 rel=99 In which year did a purchase of Alaska happen?
cosine=0.4714 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.2673 rank=3 qid=3 rel=0 1867 - U.S. President Andrew Jackson proclaims treaty for purchase of Alaska from Russia.
cosine=0.4216 rank=2 qid=3 rel=0 William Seward negotiated a purchase of Alaska for $7.2 million.
```

(2)Better tokenization algorithms

• Considering the punctuation mark

Since we only use simple white-space to partition the tokens, sometimes we may meet the following circumstances: using pid=1 as an example:

the query sentence is: Give us the name of the volcano that destroyed the ancient city of Pompeii, whose key word is **Pompeii**.

However, the four documents are as follows:

Qid=1

```
qid=1 rel=99 Give us the name of the volcano that destroyed the ancient city of Pompeii
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.
qid=1 rel=0 You can see Vesuvius in the background, near ruins of Pompei; its last eruption was in 1944.
qid=1 rel=0 Vesuvius is located near the ruins of the destroyed city of Pompeii.
qid=1 rel=0 In 79 A.D., this ancient city was buried in an avalanche of hot ash from Mount Vesuvius.
```

Since the pink-tagged words are all follows with punctuation, which will be ignored when determining the cosine-similarities. So what we need to do is to use the punctuation-ignore algorithm to refine the system.

And the related coding is as follows:

String pucfree=docText.replaceAll("[\\p{Punct}]+", " ");

Result for using punctuation-ignore method

cosine=0.2927 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=0.4282 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=0.3858 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=0.4336 rank=2 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
cosine=0.6682 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.2474 rank=2 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.4629 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=0.6529 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.8750 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
cosine=0.4743 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.3627 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
cosine=0.2236 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=0.1455 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=0.4082 rank=2 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=0.2548 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=0.2758 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=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=0.6931818181818182

Now, we can see that the MRR has increased dramatically.

Also, the following is using the qid=14 as an example to fully illustrate this stemming algorithm.The key word is “Commodores” and if we consider the punctuation, we will ignore the key word in document 1, which will cause its lower score.

Original rank for qid=14

rank= 0 qid=14 rel=99 Who was the lead singer for the Commodores
rank= 2 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
rank= 3 qid=14 rel=0 The Commodores are an American funk/soulband, which was at its peak in the 1970s.
rank= 1 qid=14 rel=0 The Commodores originally came together from groups the Mystics and the Jays.

updated rank for qid=14

rank= 0 qid=14 rel=99 Who was the lead singer for the Commodores
rank= 1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
rank= 3 qid=14 rel=0 The Commodores are an American funk/soulband, which was at its peak in the 1970s.
rank= 2 qid=14 rel=0 The Commodores originally came together from groups the Mystics and the Jays.

- **Considering error from unnecessary words(stop words, such like is, am, are, have)**

This error is caused by over consideration about the stop words(such like is, am,are,the), which has no meaning function in the sentence. For example, in this sentence,"The car has the red the girl like." The word 'the' may contain a big ratio in calculating the cosine similarity, which will cause mistakes.

So, I use the method that whenever we meet the word in the stopwords.txt, I will skip it and continue to considerate the next words.

And the related coding is as follows:

```
String stopwords = readToString("/Users/leixiao/git/hw3-leixiao2/hw3-leixiao2/src/
main/resources/stopwords.txt");
Pattern p = Pattern.compile(temp);
Matcher matcher = p.matcher(stopwords);
/** delete the same token or stopwords*/
```

if ((wordList.indexOf(temp) != -1)||matcher.find()) continue;

However, in the given database, the method could cause MRR becoming smaller, which indicates that the method might lower the performance of the system. I think it is related to the original material, but in general it will improve the performance of the system.

Result for using stopwords

cosine=0.1833 rank=3 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=0.2268 rank=3 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=0.6708 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.5661 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=0.7538 rank=1 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
cosine=0.5345 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.5000 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=0.4781 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=0.4364 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.9129 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
cosine=0.6325 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.1543 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
cosine=0.3849 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.7071 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.2649 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=0.6547 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.2500 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=0.1132 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 offer customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
cosine=0.1581 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=0.4714 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=0.646825396825397

(3) Better or different similarity measures

Using cosine similarity as the similarity calculation

cosine=0.2927 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=0.4282 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=0.3858 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=0.4336 rank=2 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
cosine=0.6682 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.2474 rank=2 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
cosine=0.4629 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=0.6529 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.8750 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
cosine=0.4743 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.3627 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
cosine=0.2236 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=0.1455 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=0.4082 rank=2 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=0.2548 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=0.2758 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=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=0.6931818181818182

• unnormalized TF-IDF

This method is to measure the overlap of the query and the document, which considers about the term frequency.

And this method involves in two aspects: tf and idf.

TF is short for term frequency, which stands for the number of times that t, which is the term in the query, occurs in each document.

DF is short for document frequency, which stands for the number of documents which contains the term t in the query.

And IDF is short for inverse document frequency, which is N/df. Actually, we use log (N/df) instead of N/df to "dampen" the effect of idf.

So, the final formula for the tf-idf is as follows.

$$\text{score(q,d)}=\sum_{t\in q} tf_t * (\log(N/df_t))$$

The related coding is as follows:

```
/**idf**/
int count=0;
int sum=0;
for (i = 0; i < relList.size(); i++) {
 if (qIdList.get(i) == id) {
            if (relList.get(i) == 99) {
                        querynumber = i;
                        count=0;
                        sum++;
                        df.add(new HashMap<String, Double>());
            }
             else count++;
 else {
```

```
 for (Map.Entry<String, Integer> entry : ((HashMap<String, Integer>)al.get(querynumber)).entrySet())
{
String a = entry.getKey();
int l=0;
for(int k=0;k<count;k++){
 if(((HashMap<String, Integer>)al.get(i-k-1)).containsKey(a)){
            l++;//ocurrency of term
}
}
double d=d=(double)count/(double)l;
double e= Math.log(d);
df.get(sum-1).put(a,e);
}
id = qIdList.get(i);
i--;
}
}

/**tf**/
for (Map.Entry<String, Integer> entry : queryVector.entrySet()){

                    String a = entry.getKey();
                    if (docVector.containsKey(a)){
                              int s=docVector.get(a);
                              tf.put(a,(double)s/(double)docVector.size());
                    }
                    else tf.put(a,0.0);
          }

/** tf-idf**/
double tfidf=0.0;
          for (Map.Entry<String, Double> entry : ddf.entrySet()) {
                    String a = entry.getKey();
                    Double b = entry.getValue();
                    if(tf.containsKey(a)){
                              tfidf+=b*tf.get(a);
                    }
```

And the results are as follows:

Using TF-IDF as the similarity calculation

df-idf=0.0561 rank=3 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.
df-idf=0.1052 rank=3 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.
df-idf=0.2003 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
df-idf=0.0300 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.
df-idf=0.1772 rank=1 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
df-idf=0.1294 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
df-idf=0.1337 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.
df-idf=0.0956 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.
df-idf=0.1310 rank=1 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
df-idf=0.4717 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
df-idf=0.0000 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
df-idf=0.0000 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
df-idf=0.0000 rank=1 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
df-idf=0.3387 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
df-idf=0.0814 rank=2 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.
df-idf=0.0541 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
df-idf=0.0246 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.
df-idf=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.
df-idf=0.0198 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.
df-idf=0.0688 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=0.7541666666666667

We can see the the result above, we might understand that this method could improve the performance of the system dramatically, which could improve the MRR from 0.67 to 0.75. However, we could notice some problems here and take the qid=11 as the example to fully understand the drawbacks of tf-idf similarity.

Qid=11

qid=11 rel=99 Where is Devil's Tower
qid=11 rel=1 Devils Tower can be found in Crook County
qid=11 rel=0 To the west, across the Wyoming border, is the staggeringly beautiful Devil's Tower National Monument.
qid=11 rel=0 Devil's Tower is an igneous intrusion that rises dramatically 1,267 feet (386 m) above the surrounding terrain.
qid=11 rel=0 In 1941, Petzoldt joined other rock climbers to rescue a marooned parachutist who had landed atop Devil's Tower

The query and documents for qid=11 are as follows:

As we can see from the graph above, we can know that from the tokenize and stemming algorithm we did before, the terms we stem from the query is only “devils” and “tower”, and according to the tf-idf method, when we calculate the df for the two words, we will get that df(devils)=4, and df(tower)=4. And then when we calculate idf(devils)=log(4/df)=0.0, and idf(devils)=log(4/df)=0.0.

Then to get the final score, we get tf-idf=0.0*tf=0.0, which will cause for every document sentence, the score is 0.

And when we do the ranking, the related document, which relevance value is 1, will absolutely have rank=1, which will cause the increase of MRR.

And when we do not notice that phenomenon, we will have a wrong impression on this method, since the tf-idf cannot improve the performance so much.

However, this method did do much better than the cosine similarity because of its consideration about the query and document separately.

As we all know that if the query term is very rare in the whole collection, then the document which has this term could have more possibilities to be the most relevant one, and if the query term is very frequent in the whole collection, then the document which has the most frequency for this term could be more likely to be the most relevant one than others.

So, considering this, the tf-idf has been developed, which not only consider the term frequency, but also the document frequency, which will have a better performance than the cosine similarity that only consider the term frequency.

However, there is another formula to calculate the score of tf-idf:

$$w_{t,d} = \log(1 + \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

And if we use this formula , we will get a better performance for this system.

In this case, we still have two conditions since we have two different understanding about tf.

So, if we understand tf as the occurrence of the term in document d, the results will be as follows:

Using tf-idf as the similarity calculation

df-idf=0.6799 rank=3 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.
df-idf=1.6408 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.
df-idf=0.7615 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
df-idf=0.2810 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.
df-idf=1.0425 rank=1 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
df-idf=0.7615 rank=1 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
df-idf=0.8793 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.
df-idf=0.9609 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.
df-idf=0.6799 rank=1 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
df-idf=2.1212 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
df-idf=0.0000 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
df-idf=0.0000 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
df-idf=0.0000 rank=1 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
df-idf=1.5230 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
df-idf=0.7615 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.
df-idf=0.2810 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
df-idf=0.2810 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.
df-idf=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.
df-idf=0.2810 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.
df-idf=0.4528 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=0.8625

So, if we understand tf as the occurrence of the term in document d/the total number of terms in document d, the results will be as follows:(which is the same as the first formula)

Using TF-IDF as the similarity calculation

df-idf=0.0561 rank=3 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.
df-idf=0.1052 rank=3 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.
df-idf=0.2003 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
df-idf=0.0300 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.
df-idf=0.1772 rank=1 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
df-idf=0.1294 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
df-idf=0.1337 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.
df-idf=0.0956 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.
df-idf=0.1310 rank=1 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
df-idf=0.4717 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
df-idf=0.0000 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
df-idf=0.0000 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
df-idf=0.0000 rank=1 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
df-idf=0.3387 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
df-idf=0.0814 rank=2 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.
df-idf=0.0541 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
df-idf=0.0246 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.
df-idf=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.
df-idf=0.0198 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.
df-idf=0.0688 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=0.7541666666666667

I think the choice of the formula should be associated with the database, with different cases, the formula will have different influence on the performance.

• the Dice coefficient

This method is comely used to measure the overlap of the two sentences and the method does not consider about the term frequency, but only consider about the the number of the words which exists in both sentences, and the total words in the each sentence.

The formula of calculating the score of the Dice coefficient is as follows:

$dice(A,B) = 2*|A \cap B| / (|A| + |B|)$

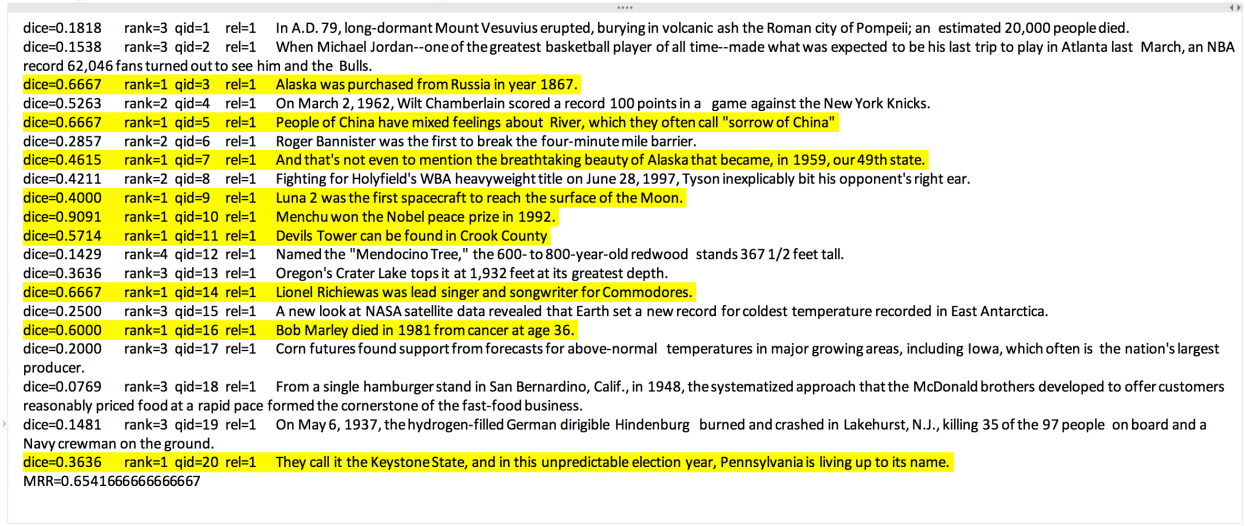
And in this case, the score of the Dice coefficient is always between 0 and 1.

The related coding is as follows:

```
/**dice score**/
int querysize = queryVector.size();
int documentsize = docVector.size();
int m=querysize+documentsize;
for (Map.Entry<String, Integer> entry : queryVector.entrySet()){
        String a = entry.getKey();
        if (docVector.containsKey(a)){
                p++;
        }
}
dice_coefficient = (double) (2*p) / (double) (m);
```

The result is as follows:

Using dice coefficient as the similarity calculation



```
dice=0.1818    rank=3  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.
dice=0.1538    rank=3  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.
dice=0.6667    rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
dice=0.5263    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.
dice=0.6667    rank=1  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
dice=0.2857    rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
dice=0.4615    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.
dice=0.4211    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.
dice=0.4000    rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
dice=0.9091    rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
dice=0.5714    rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
dice=0.1429    rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood  stands 367 1/2 feet tall.
dice=0.3636    rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
dice=0.6667    rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
dice=0.2500    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.
dice=0.6000    rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
dice=0.2000    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.
dice=0.0769    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 offer customers
reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
dice=0.1481    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.
dice=0.3636    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=0.6541666666666667
```

We can see from the result above, although the score of each sentence is different, the rank of each sentence is the same. The final result of MRR is almost the same to the database, since they all select 8 most relevant document to 20 queries.

• the Jaccard coefficient

This method is comely used to measure the overlap of the two sentences and the method does not consider about the term frequency, but only consider about the the number of the words which exists in both sentences, and the total words in the two sentences.

The formula of calculating the score of the Jaccard coefficient is as follows:

$jaccard(A,B) = |A \cap B| / |A \cup B|$

$jaccard(A,A) = 1$

$jaccard(A,B) = 0 \text{ if } A \cap B = 0$

And in this case, the score of the Jaccard coefficient is always between 0 and 1.

The related coding is as follows:

```
/**jaccard score**/
Map<String, Integer> temp = new HashMap<String, Integer>();
temp.putAll(queryVector);
temp.putAll(docVector);
int m=temp.size();
int p=0;
for (Map.Entry<String, Integer> entry : queryVector.entrySet()){
        String a = entry.getKey();
        if (docVector.containsKey(a)){
                p++;
        }
}
jaccard_coefficient = (double) (p) / (double) (m);
```

The result is as follows:

Using jaccard coefficient as the similarity calculation

```
jaccard=0.1000 rank=3  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.
jaccard=0.0833 rank=3  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.
jaccard=0.5000 rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
jaccard=0.3571 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.
jaccard=0.5000 rank=1  qid=5   rel=1   People of China have mixed feelings about  River, which they often call "sorrow of China"
jaccard=0.1667 rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barrier.
jaccard=0.3000 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.
jaccard=0.2667 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.
jaccard=0.2500 rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
jaccard=0.8333 rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
jaccard=0.4000 rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
jaccard=0.0769 rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood  stands 367 1/2 feet tall.
jaccard=0.2222 rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
jaccard=0.5000 rank=1  qid=14  rel=1   Lionel Richiewas was lead  singer and songwriter for Commodores.
jaccard=0.1429 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.
jaccard=0.4286 rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
jaccard=0.1111 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.
jaccard=0.0400 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 offer
customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
jaccard=0.0800 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.
jaccard=0.2222 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=0.6541666666666667
```

We can see from the result above, although the score of each sentence is different, the rank of each sentence is the same. The final result of MRR is the almost the same to the database, since they all select 8 most relevant document to 20 queries.

• BM25

The basic idea of BM25 is just like the tf-idf, which is considering about the query terms' occurrence in the document, but it do not consider about the term frequency within each document.

The basic formula is as follows:

Given a query Q, containing keywords $q_1, ..., q_n$, the BM25 score of a document D is:

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

and f(qi, D) means the term frequency of qi in D, k1 can be valued from 1.2 to 2.0, and b is 0.75.

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

And in this formula, n(qi) is the number of documents containing the term qi, and N is the total number of documents in the collection.

And the results are as follows:

Using BM25 as the similarity calculation

BM25=0.1181 rank=3 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.
BM25=0.1409 rank=3 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.
BM25=1.0449 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
BM25=0.5132 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.
BM25=0.7853 rank=1 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
BM25=0.8742 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
BM25=0.4449 rank=2 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
BM25=0.4244 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.
BM25=0.5141 rank=1 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
BM25=1.6837 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
BM25=0.8168 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
BM25=0.0849 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
BM25=0.2431 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
BM25=0.7185 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
BM25=0.1479 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.
BM25=0.7595 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
BM25=0.1399 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.
BM25=0.0425 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 offer customers reasonably priced food at a rapid pace formed the cornerstone of the fast-food business.
BM25=0.1197 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.
BM25=0.2933 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=0.6041666666666667