HW3 Report

Andrew ID : yichenca

Author: Cai Yichen

Date: 2014-10-19

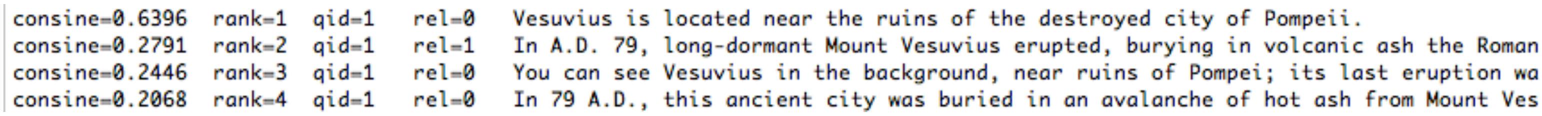
Task 1

Error Analysis

In this task, there are 20 queries in total. By observing the result output, only one relevant answer is retrieved according to our naive white space tokenization with cosine similarity algorithm. Namely, the correctness rate is only 1/20, which is not acceptable. Below is the error analysis of some selective queries.

Q1: Give us the name of the volcanoes that destroyed the ancient city of Pompeii

The output is :

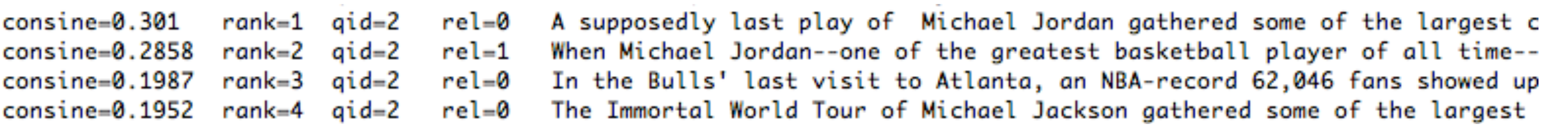


```
consine=0.6396  rank=1  qid=1   rel=0   Vesuvius is located near the ruins of the destroyed city of Pompeii.
consine=0.2791  rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman
consine=0.2446  rank=3  qid=1   rel=0   You can see Vesuvius in the background, near ruins of Pompei; its last eruption wa
consine=0.2068  rank=4  qid=1   rel=0   In 79 A.D., this ancient city was buried in an avalanche of hot ash from Mount Ves
```

The relevant answer ranks 2nd with cosine value 0.2791. We can tell the match words are city, of , the. The word Pompeii; doesn't match Pompeii in query, because of the punctuation. Also the word volcanic can't match volcanoes in query. In addition, the relevant answer has too many useless and redundant words, which decrease its cosine value.

Q2: What has been the largest crowd to ever come see Michael Jordan

The output is :

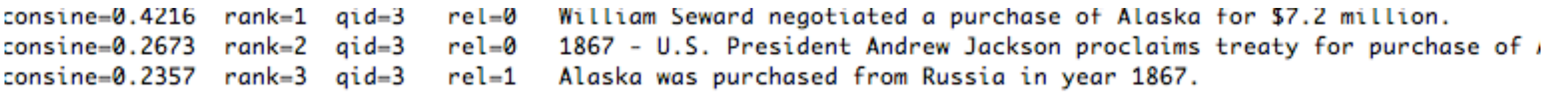


```
consine=0.301   rank=1  qid=2   rel=0   A supposedly last play of  Michael Jordan gathered some of the largest c
consine=0.2858  rank=2  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all time--
consine=0.1987  rank=3  qid=2   rel=0   In the Bulls' last visit to Atlanta, an NBA-record 62,046 fans showed up
consine=0.1952  rank=4  qid=2   rel=0   The Immortal World Tour of Michael Jackson gathered some of the largest
```

The relevant answer ranks 2nd with cosine value 0.2858. We can tell word Jordan—one is not disjoint so that it cannot math the word Jordan in Query. In addition, the relevant answer has too many useless and redundant words, which decrease its cosine value.

Q3: In which year did a purchase of Alaska happen?

The output is :

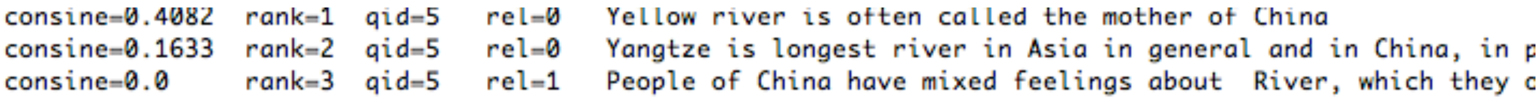


```
consine=0.4216  rank=1  qid=3   rel=0   William Seward negotiated a purchase of Alaska for $7.2 million.
consine=0.2673  rank=2  qid=3   rel=0   1867 - U.S. President Andrew Jackson proclaims treaty for purchase of
consine=0.2357  rank=3  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
```

The relevant answer ranks 3rd with cosine value 0.2357. The word purchased is not match the word purchase in query cause the different tense.

Q5: What river is called China's Sorrow?

The output is :

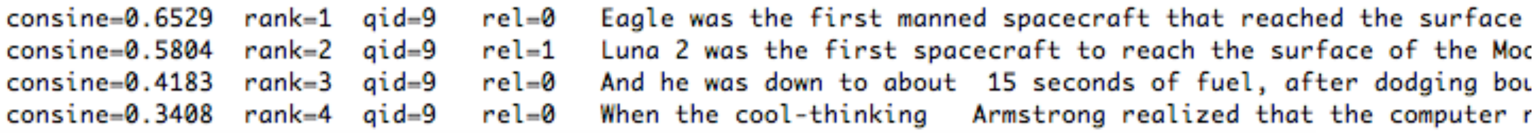


```
consine=0.4082  rank=1  qid=5   rel=0   Yellow river is often called the mother of China
consine=0.1633  rank=2  qid=5   rel=0   Yangtze is longest river in Asia in general and in China, in p
consine=0.0     rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which they c
```

The relevant answer ranks 3rd with cosine value 0.0. The word River, doesn't match river because of the capital case 'R' and punctuation ','. And 'China's Sorrow?' can't match ' "sorrow of China" ' . So if we omit the possessive case ' 's ', and punctuation ' ? ', it can increase the correctness rate.

Q9: What was the first spaceship on the moon

The output is:

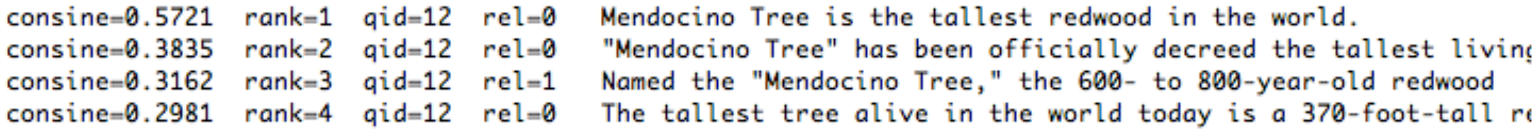


```
consine=0.6529  rank=1  qid=9   rel=0   Eagle was the first manned spacecraft that reached the surface
consine=0.5804  rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moo
consine=0.4183  rank=3  qid=9   rel=0   And he was down to about  15 seconds of fuel, after dodging bou
consine=0.3408  rank=4  qid=9   rel=0   When the cool-thinking   Armstrong realized that the computer r
```

The relevant answer ranks 2nd with cosine value 0.5804. First, the word ' Moon' doesn't match 'moon' because of capital case 'M'. Plus, Luna 2 actually is a word represent the spaceship, but it is split wrongly. In addition, the word 'spacecraft' is synonymous with the word 'spaceship'. If we can recognize synonyms, it will increase cosine value.

Q12: What is the height of the tallest redwood?

The output is:

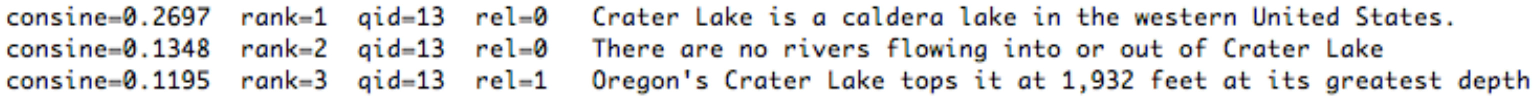


```
consine=0.5721  rank=1  qid=12  rel=0   Mendocino Tree is the tallest redwood in the world.
consine=0.3835  rank=2  qid=12  rel=0   "Mendocino Tree" has been officially decreed the tallest living
consine=0.3162  rank=3  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood
consine=0.2981  rank=4  qid=12  rel=0   The tallest tree alive in the world today is a 370-foot-tall r
```

The relevant answer ranks 3rd with cosine value 0.3162. The word "tall" can't match "tallest" because of comparative degree.

Q13: How deep is Crater Lake?

The output is:

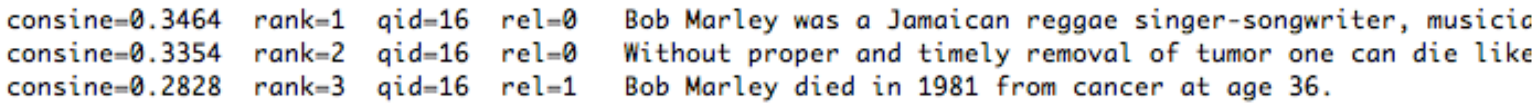


```
consine=0.2697  rank=1  qid=13  rel=0   Crater Lake is a caldera lake in the western United States.
consine=0.1348  rank=2  qid=13  rel=0   There are no rivers flowing into or out of Crater Lake
consine=0.1195  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth
```

The relevant answer ranks 3rd with cosine value only 0.1195. The word ‘depth’ doesn’t match ‘deep’, because of the word property. Actually, the two words tell the same thing.

Q16: When did Bob Marley die

The output is :



```
consine=0.3464  rank=1  qid=16  rel=0   Bob Marley was a Jamaican reggae singer-songwriter, musicia
consine=0.3354  rank=2  qid=16  rel=0   Without proper and timely removal of tumor one can die like
consine=0.2828  rank=3  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
```

The relevant answer ranks 3rd with cosine vale 0.2828. First, because of different tense, word ‘died’ doesn’t match ‘die’. Plus, the word ‘did’ from the 1st answer does match the word ‘did’ in query, which increase its cosine value wrongly. So it is a useless stop word.

Summary :

Error Type	Queries	Examples
There are too many useless stop words, which decrease the cosine value.	Q1,Q2,Q4,Q8,Q16	an, in, of, the , on, etc.
Different tenses, the stemming problem	Q3,Q7,Q8,Q16	die, died; become, became; etc.
Punctuation , such comma, hyphen	Q6,Q18,Q19,Q14,Q20,Q1,Q5	? , - “ “ ; etc
Capital case and lower case	Q5,Q9,Q15,Q17	R’ ‘r’ ; ‘M’ ‘m’
word property	Q12,Q13	‘deep, depth’ ‘tall, tallest’
possessive case	Q5,Q11	Devil, Devil’s
synonyms	Q9	spaceship, spacecraft
plural nouns	Q6	minute, minutes
redundancy and irreverent information	Q1, Q2, Q4	refer to specific answers

In task 2, I try to use different tokenization methods, such as delete punctuation and stop words, omit capital case; and different stemming methods, such as StanfordLemmatizers ; and different similarity methods, such as TF - IDF.

Task 2

System Design

Better Tokenization algorithm:

In task 2, based on the error analysis of Task1, the errors such as capitalization, punctuation, too many stop words would impair our cosine similarity. So. I try to implement better tokenization algorithm, which combining different tokenization methods including naive white space tokenizer, capitalization deletion, punctuation deletion, stop-word deletion. These methods are all in Task2Tokenization.java.

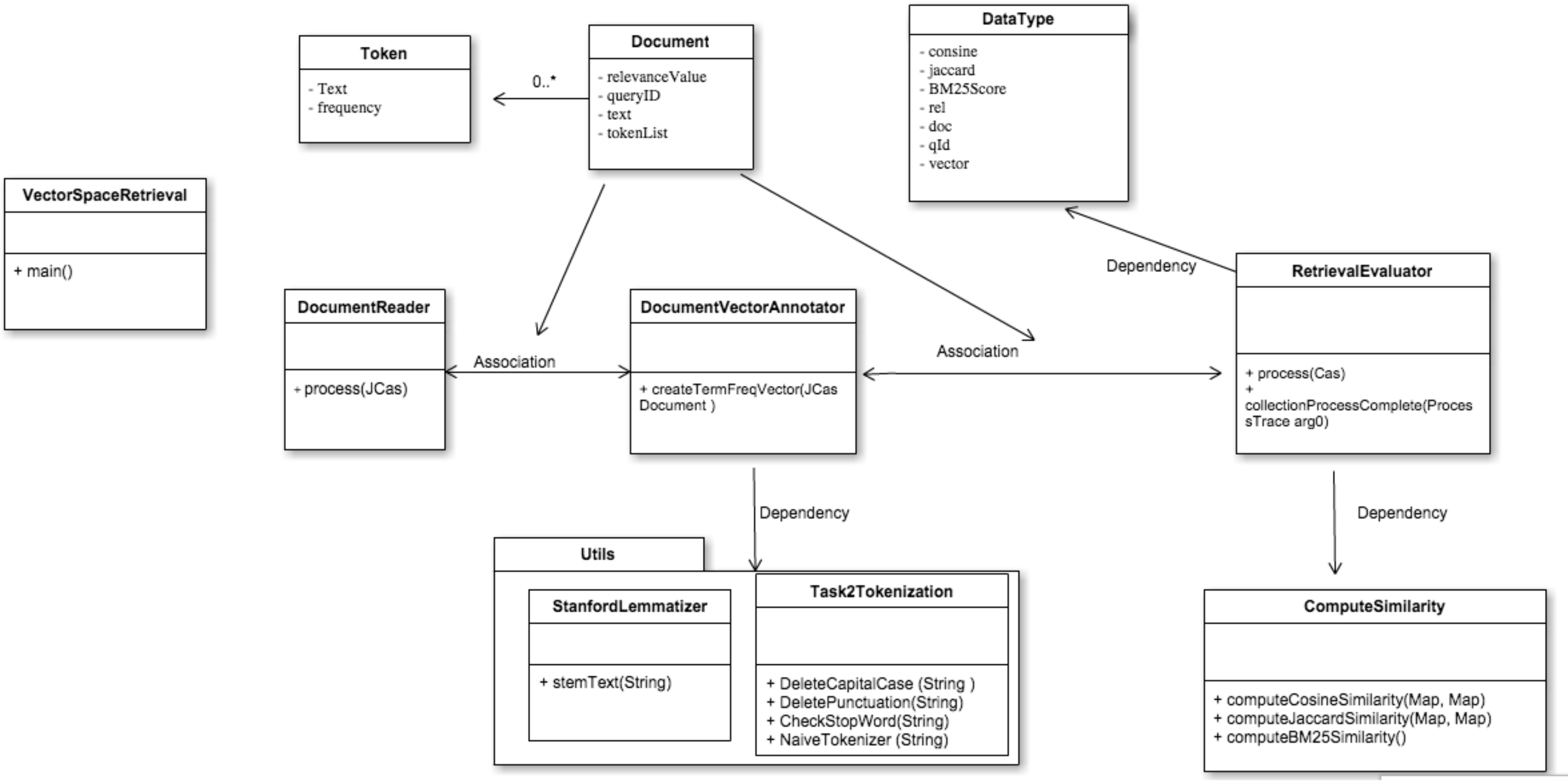
Better Stemming algorithm:

Because the simple vector model does not understand that die and died denote the same thing in most contexts. I use the Stanford lemmatizer provided in util.StanfordLemmatizer to lemmatize these kind of words.

Different similarity measures:

Basically, I implement another 2 different methods : Jaccard Coefficient and BM25. All these methods are in ComputeSimilarity.java.

UML Diagram:



System Description

As same as task1, task2 system consists of VectorSpaceRetrieval, DocumentReader, DocumentVectorAnnotator and RetrievalEvaluator. The DataType class contains the useful information (cosine, jaccard, BM25Score, rel, doc, qId, vector) for RetrievalEvaluator.

I include all the tokenization methods in Task2Tokenization class :

1. DeleteCapitalCase(String) : this method converts all the capitalized character to lowercase ;
2. DeletePunctuation(String) : this method deletes all the punctuations in a given sentence ;
3. NaiveTokenizer(String) : this method just uses the provided tokenizer0 to spit sentences by white space;
4. CheckStopWord(String) : this method reads content from stopwords.txt and store all the words in Set<String> as dictionary, then it checks if a given string is in this dictionary. If it exists, return true, else return false;

In addition, I use StanfordLemmatizer to lemmatize the words with same meaning so that it will increase cosine similarity.

In the similarity measures part, I put all the similarity methods in ComputeSimilarity class.

1. ComputeCosineSimilarity(Map query, Map document) : this method compute the cosine value between two vectors, say query and document.

The formula is :

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

2. ComputeJaccardSimilarity(Map query, Map document) : this method measures the similarity between two sets, the value is the size of the intersection divided by the size of the union of the same sets. And I use a similar method to implement Jaccard.

The formular is :

$$J(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)},$$

3. ComputeBM25Similarity( ) : this method bases on the query teams appearing each document, regardless of the inter-relationship between the query terms within a document.

The formular is :

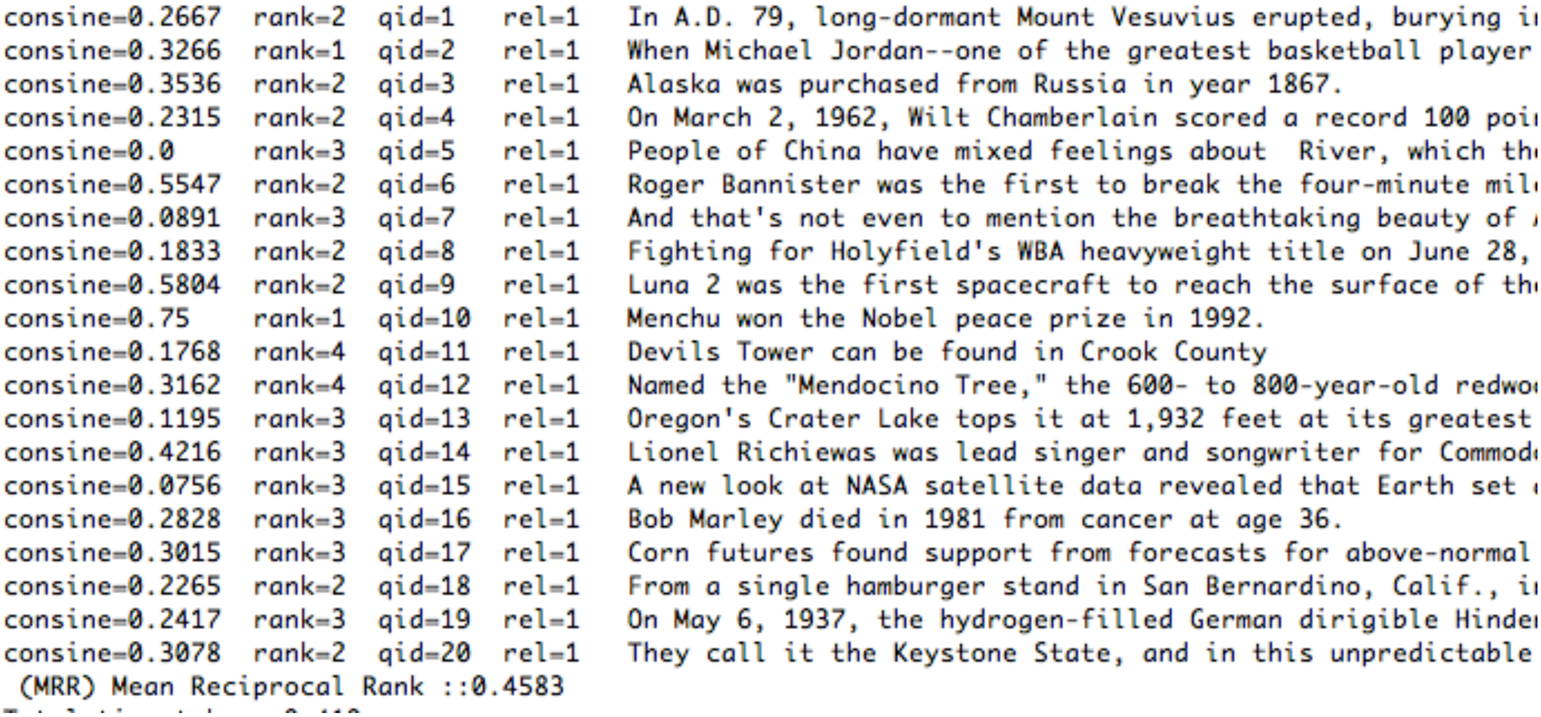
$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

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

After testing , I choose k1 = 1.2 and b =0.75 as the parameter for this BM25 algorithm.

Comparison Testing:

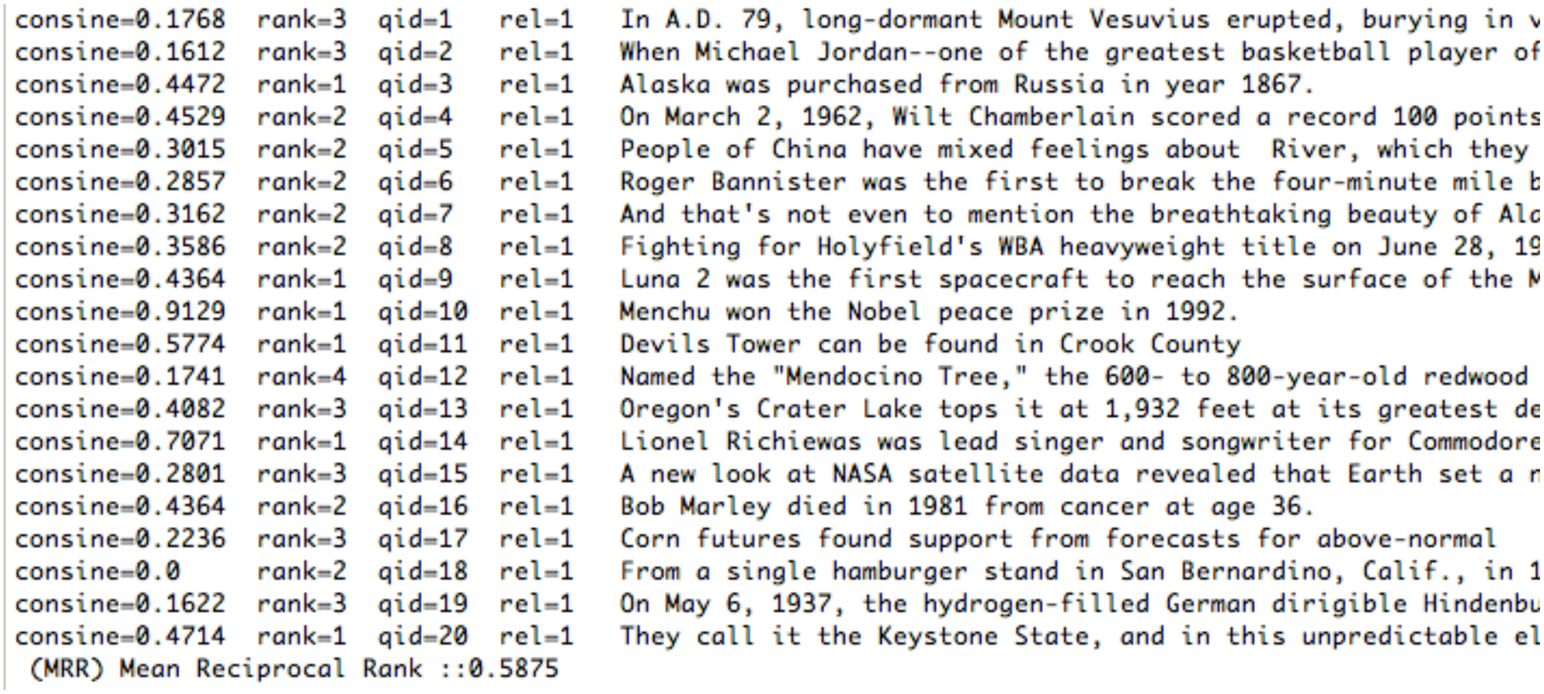
1. Use DeleteCapitalization and White space tokenizer with cosine similarity, the result shows below:



```
consine=0.2667  rank=2  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying i
consine=0.3266  rank=1  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player
consine=0.3536  rank=2  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
consine=0.2315  rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 poi
consine=0.0     rank=3  qid=5   rel=1   People of China have mixed feelings about  River, which th
consine=0.5547  rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mil
consine=0.0891  rank=3  qid=7   rel=1   And that's not even to mention the breathtaking beauty of
consine=0.1833  rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28,
consine=0.5804  rank=2  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of th
consine=0.75    rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
consine=0.1768  rank=4  qid=11  rel=1   Devils Tower can be found in Crook County
consine=0.3162  rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwo
consine=0.1195  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest
consine=0.4216  rank=3  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commod
consine=0.0756  rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set
consine=0.2828  rank=3  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
consine=0.3015  rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal
consine=0.2265  rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., i
consine=0.2417  rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hinde
consine=0.3078  rank=2  qid=20  rel=1   They call it the Keystone State, and in this unpredictable
 (MRR) Mean Reciprocal Rank ::0.4583
```

We can see the MRR increases to 0.4583 from 0.4375. So this tokenizer does improve the system performance.

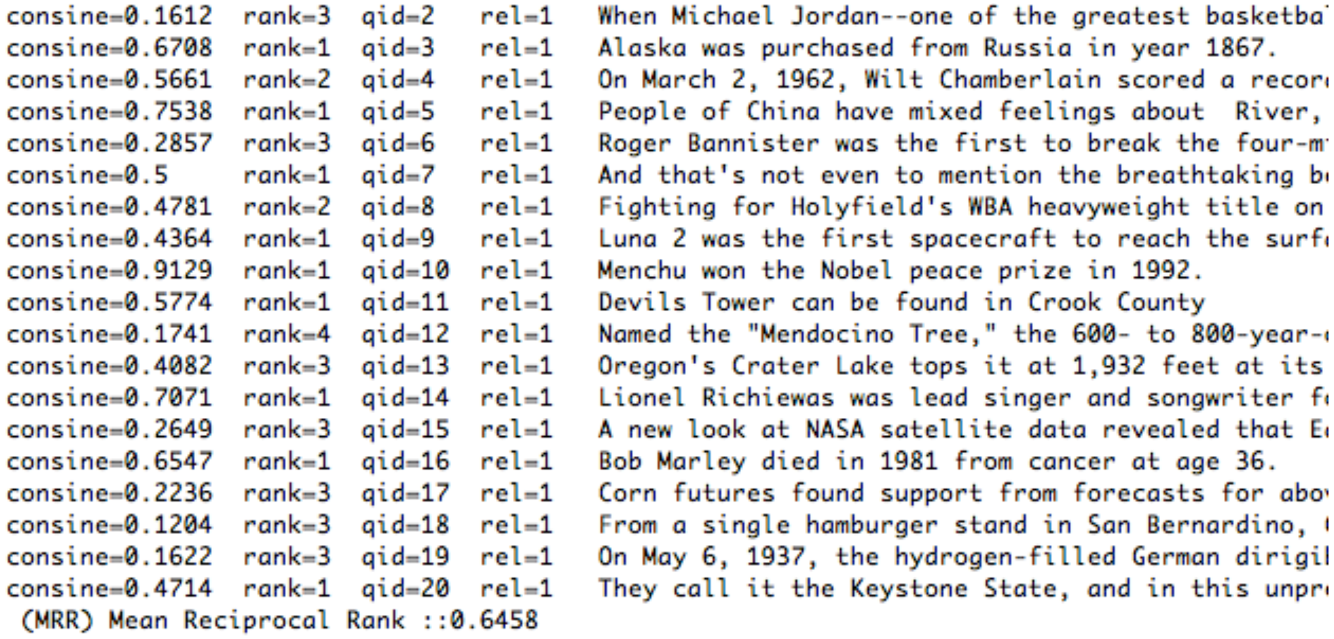
2. Use DeleteCapitalization ,DeletePunctuation, CheckStopwords and White space tokenizer with cosine similarity, the result shows below:



```
consine=0.1768  rank=3  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in v
consine=0.1612  rank=3  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of
consine=0.4472  rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
consine=0.4529  rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points
consine=0.3015  rank=2  qid=5   rel=1   People of China have mixed feelings about  River, which they
consine=0.2857  rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile b
consine=0.3162  rank=2  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alc
consine=0.3586  rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 19
consine=0.4364  rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the M
consine=0.9129  rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
consine=0.5774  rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
consine=0.1741  rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood
consine=0.4082  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest de
consine=0.7071  rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodore
consine=0.2801  rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set a r
consine=0.4364  rank=2  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
consine=0.2236  rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal
consine=0.0     rank=2  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1
consine=0.1622  rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenbu
consine=0.4714  rank=1  qid=20  rel=1   They call it the Keystone State, and in this unpredictable el
 (MRR) Mean Reciprocal Rank ::0.5875
```

We can see the MRR further increase by about 0.13. So combining these tokenization methods will improve the system performance well.

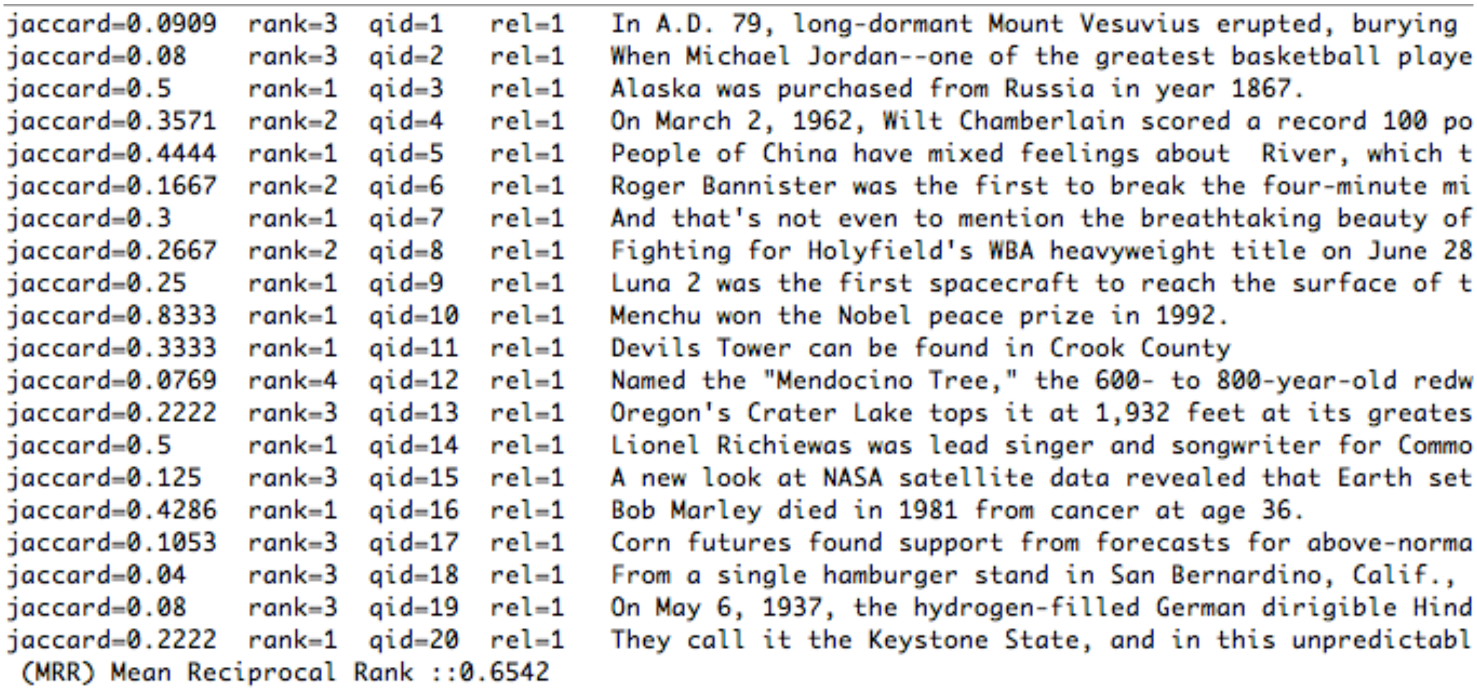
3. Then we add StanfordLemmatizer method, the result shows below:



```
consine=0.1612  rank=3  qid=2   rel=1   When Michael Jordan--one of the greatest basketba
consine=0.6708  rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
consine=0.5661  rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a recor
consine=0.7538  rank=1  qid=5   rel=1   People of China have mixed feelings about  River,
consine=0.2857  rank=3  qid=6   rel=1   Roger Bannister was the first to break the four-m
consine=0.5     rank=1  qid=7   rel=1   And that's not even to mention the breathtaking b
consine=0.4781  rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on
consine=0.4364  rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surf
consine=0.9129  rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
consine=0.5774  rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
consine=0.1741  rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-
consine=0.4082  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its
consine=0.7071  rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter f
consine=0.2649  rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that E
consine=0.6547  rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
consine=0.2236  rank=3  qid=17  rel=1   Corn futures found support from forecasts for abo
consine=0.1204  rank=3  qid=18  rel=1   From a single hamburger stand in San Bernardino,
consine=0.1622  rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigi
consine=0.4714  rank=1  qid=20  rel=1   They call it the Keystone State, and in this unpr
 (MRR) Mean Reciprocal Rank ::0.6458
```

We can see the MRR increases by about 0.6. Here, we have combined all the implemented tokenizers and stemmer.

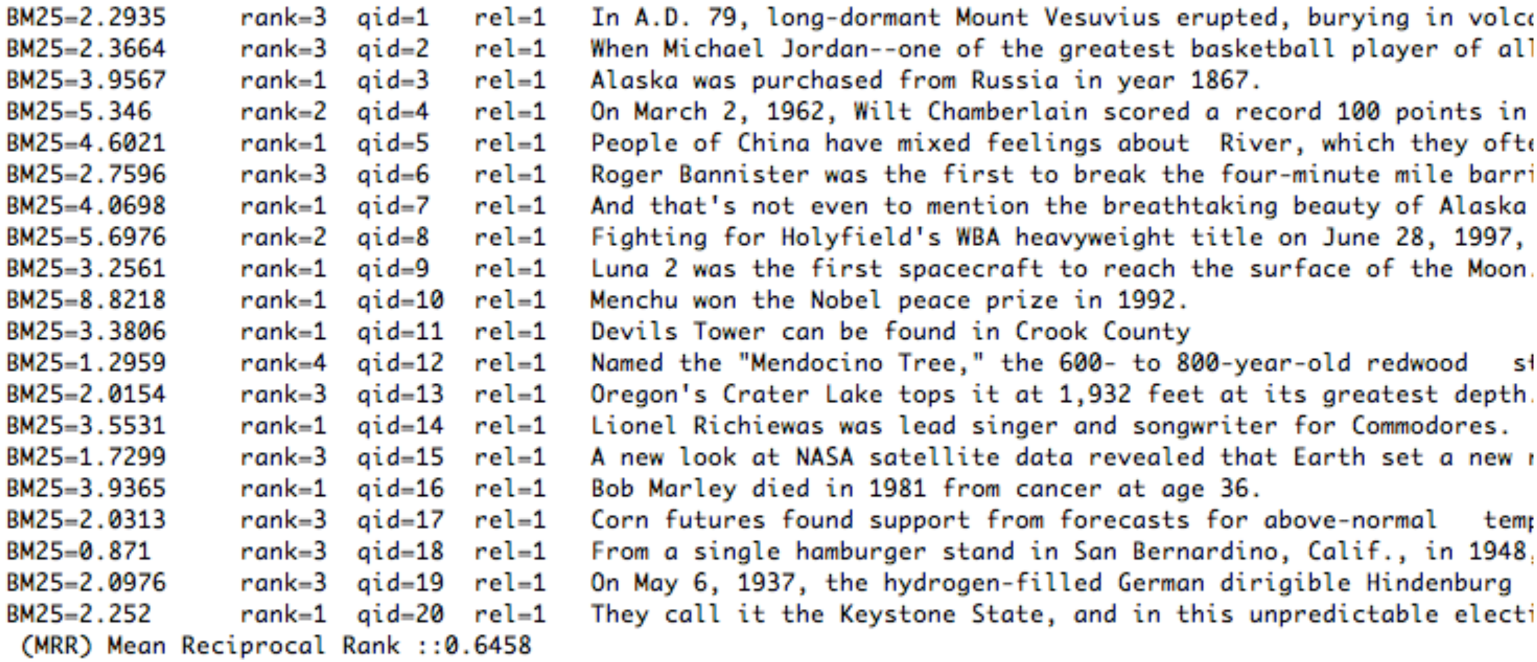
4. Try to use Jaccard coefficient , the result shows below:



```
jaccard=0.0909  rank=3  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying
jaccard=0.08    rank=3  qid=2   rel=1   When Michael Jordan--one of the greatest basketball playe
jaccard=0.5     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 po
jaccard=0.4444  rank=1  qid=5   rel=1   People of China have mixed feelings about  River, which t
jaccard=0.1667  rank=2  qid=6   rel=1   Roger Bannister was the first to break the four-minute mi
jaccard=0.3     rank=1  qid=7   rel=1   And that's not even to mention the breathtaking beauty of
jaccard=0.2667  rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28
jaccard=0.25    rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of t
jaccard=0.8333  rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
jaccard=0.3333  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 redw
jaccard=0.2222  rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greates
jaccard=0.5     rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commo
jaccard=0.125   rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set
jaccard=0.4286  rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
jaccard=0.1053  rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-norma
jaccard=0.04    rank=3  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif.,
jaccard=0.08    rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hind
jaccard=0.2222  rank=1  qid=20  rel=1   They call it the Keystone State, and in this unpredictabl
 (MRR) Mean Reciprocal Rank ::0.6542
```

The final MRR is 0.6542. It is very similar to cosine similarity. This method chooses 9 relevant document correctly, which is same as cosine similarity.

5. Try to use BM25, the result shows below:

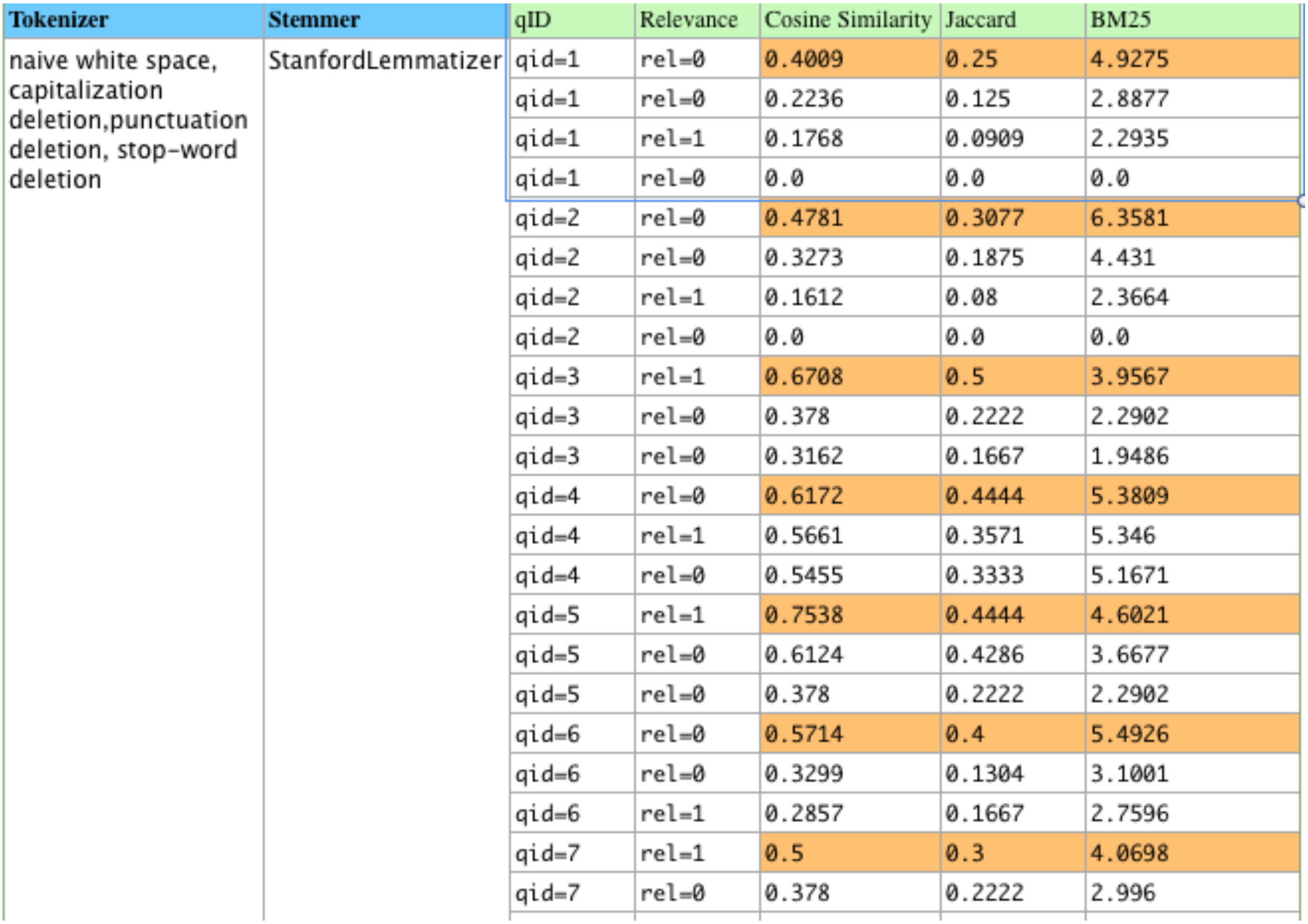


```
BM25=2.2935     rank=3  qid=1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volc
BM25=2.3664     rank=3  qid=2   rel=1   When Michael Jordan--one of the greatest basketball player of all
BM25=3.9567     rank=1  qid=3   rel=1   Alaska was purchased from Russia in year 1867.
BM25=5.346      rank=2  qid=4   rel=1   On March 2, 1962, Wilt Chamberlain scored a record 100 points in
BM25=4.6021     rank=1  qid=5   rel=1   People of China have mixed feelings about  River, which they ofte
BM25=2.7596     rank=3  qid=6   rel=1   Roger Bannister was the first to break the four-minute mile barri
BM25=4.0698     rank=1  qid=7   rel=1   And that's not even to mention the breathtaking beauty of Alaska
BM25=5.6976     rank=2  qid=8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997,
BM25=3.2561     rank=1  qid=9   rel=1   Luna 2 was the first spacecraft to reach the surface of the Moon.
BM25=8.8218     rank=1  qid=10  rel=1   Menchu won the Nobel peace prize in 1992.
BM25=3.3806     rank=1  qid=11  rel=1   Devils Tower can be found in Crook County
BM25=1.2959     rank=4  qid=12  rel=1   Named the "Mendocino Tree," the 600- to 800-year-old redwood   st
BM25=2.0154     rank=3  qid=13  rel=1   Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
BM25=3.5531     rank=1  qid=14  rel=1   Lionel Richiewas was lead singer and songwriter for Commodores.
BM25=1.7299     rank=3  qid=15  rel=1   A new look at NASA satellite data revealed that Earth set a new r
BM25=3.9365     rank=1  qid=16  rel=1   Bob Marley died in 1981 from cancer at age 36.
BM25=2.0313     rank=3  qid=17  rel=1   Corn futures found support from forecasts for above-normal   temp
BM25=0.871      rank=3  qid=18  rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948,
BM25=2.0976     rank=3  qid=19  rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg
BM25=2.252      rank=1  qid=20  rel=1   They call it the Keystone State, and in this unpredictable electi
 (MRR) Mean Reciprocal Rank ::0.6458
```

According to my implemented BM25 ,we can see the ranking is the same as cosine similarity. So the MRR is also the same. I’ m a little curious that the ranking of documents is all the same. Maybe the sample set is too small.

So based on these 3 tests result, the performances of the three methods, cosine similarity, Jaccard coefficient, BM25 are very similar.

Below is the part of the results. I put the detailed summary results in report.xls.



Tokenizer	Stemmer	qID	Relevance	Cosine Similarity	Jaccard	BM25
naive white space, capitalization deletion,punctuation deletion, stop-word deletion	StanfordLemmatizer	qid=1	rel=0	0.4009	0.25	4.9275
		qid=1	rel=0	0.2236	0.125	2.8877
		qid=1	rel=1	0.1768	0.0909	2.2935
		qid=1	rel=0	0.0	0.0	0.0
		qid=2	rel=0	0.4781	0.3077	6.3581
		qid=2	rel=0	0.3273	0.1875	4.431
		qid=2	rel=1	0.1612	0.08	2.3664
		qid=2	rel=0	0.0	0.0	0.0
		qid=3	rel=1	0.6708	0.5	3.9567
		qid=3	rel=0	0.378	0.2222	2.2902
		qid=3	rel=0	0.3162	0.1667	1.9486
		qid=4	rel=0	0.6172	0.4444	5.3809
		qid=4	rel=1	0.5661	0.3571	5.346
		qid=4	rel=0	0.5455	0.3333	5.1671
		qid=5	rel=1	0.7538	0.4444	4.6021
		qid=5	rel=0	0.6124	0.4286	3.6677
		qid=5	rel=0	0.378	0.2222	2.2902
		qid=6	rel=0	0.5714	0.4	5.4926
		qid=6	rel=0	0.3299	0.1304	3.1001
		qid=6	rel=1	0.2857	0.1667	2.7596
		qid=7	rel=1	0.5	0.3	4.0698
		qid=7	rel=0	0.378	0.2222	2.996