

# FINAL REPORT

Name: Ningna Wang  
AndrewID: ningnaw  
Date: 10/21/2014

## 1. Design Aspect

---

### *Building Vector space Retrieval Model*

In this task, I implemented a simple vector space retrieval system. I followed the pipeline offered by instruction document.

#### [1.1 Collection Reader](#)

The *DocumentReader* reads documents (small documents that each of which consists only one sentence) from CAS. Then, it recognizes query ID, relevance assessment and document text of each document and saves them into Document type.

#### [1.2 Annotator](#)

The *DocumentVectorAnnotator* uses a white-space tokenizer to split terms in each document. Then, I use hash map to calculate frequency of each term. And update the *tokenList* in CAS.

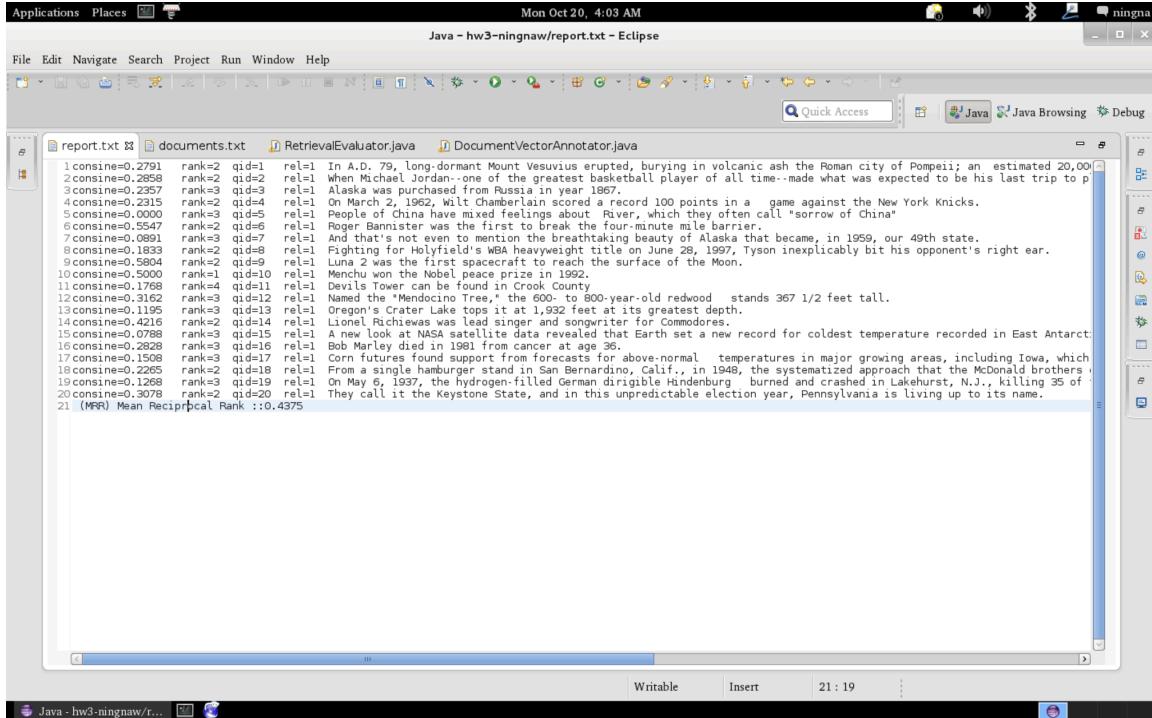
#### [1.3 Cas Consumer](#)

The *RetrievalEvaluator* is the most important part that I spent more time on implementing it. What we want to show in our output is the rank of cosine similarity of each relevant document in every query.

1. So I use hash map to store term-frequency vector. For all queries and relevant document, I use map to store their ID as well as term-frequency vector. For documents that are not relevant, I use arraylist to store all non-relevant documents of one query, then use hash map to store every query ID as well as its arraylist.
2. After calculating cosine similarity of relevant and non-relevant documents of each query, I use hash map to store query ID and arraylist that stores cosine similarities. Then sorting these values using *Collections.sort()* and save the rank of relevant documents into a new arraylist.
3. Last, I calculate MMR (Mean Reciprocal Rank) to evaluate its performance.

## 2. Error Analysis & System Improvement

In task1, I use a white-space tokenizer and cosine similarity to implement this code pipeline. Figure 1 shows the result of task 1. We can see that MMR is 0.4375, which is not that satisfying. So I consider many factors that may cause this low MMR.



```
Mon Oct 20, 4:03 AM
Java - hw3-ningnaw/report.txt - Eclipse
File Edit Navigate Search Project Run Window Help
report.txt documents.txt RetrievalEvaluator.java DocumentVectorAnnotator.java
1 consine=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. Michael Jordan is one of the greatest basketball player of all time--made what was expected to be his last trip to p
2 consine=0.2258 rank=2 qid=2 rel=1 Bob McAdoo, a forward from the University of Alaska, was purchased from Puerto Rico by the New York Knicks.
3 consine=0.2257 rank=2 qid=3 rel=1 On March 2, 1962, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
4 consine=0.2215 rank=2 qid=4 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
5 consine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China".
6 consine=0.5547 rank=2 qid=6 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
7 consine=0.0891 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.
8 consine=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.
9 consine=0.0000 rank=2 qid=9 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakehurst, N.J., killing 35 of 36 passengers and crew.
10 consine=0.5000 rank=1 qid=10 rel=1 Mandy Van der Hoeven, a Dutch singer, in 1990.
11 consine=0.1768 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County.
12 consine=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.
13 consine=0.1195 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 consine=0.4216 rank=2 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 consine=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.
16 consine=0.0000 rank=3 qid=16 rel=1 Bob Marley, the reggae legend, died in 1981.
17 consine=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
18 consine=0.2265 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematic approach that the McDonald brothers
19 consine=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 36 passengers and crew.
20 consine=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.
21 (MMR) Mean Reciprcal Rank ::0.4375
```

Figure 1. Original result

### 2.1 Bad tokenization algorithm

The tokenization algorithm I used in task 1 split sentence only by white space. This is not good since punctuations and symbols may not take into consideration. Here are examples that white-space tokenizer do not perform well.

qid	rel	error	correct
1	1	Pompeii;	Pompeii
2	1	time--made	time made
5	1	"sorrow	sorrow

Table 1.

In this case, tokening method should be improved. I use regular expression to improve tokenization algorithm, which take into account punctuations. Thus, sentences in document will be split on not only white space but also punctuations. Figure 2 shows the result of this improved algorithm. The MMR became 0.5375 from 0.4375(original result).

```

1 consine=0.3046 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.
2 consine=0.3046 rank=2 qid=2 rel=1 When Michael Jordan was one of the greatest basketball player of all time - made what was expected to be his last trip to play.
3 consine=0.2357 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 consine=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.
5 consine=0.1734 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China".
6 consine=0.6013 rank=1 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7 consine=0.1650 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.
8 consine=0.3858 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.
9 consine=0.3046 rank=1 qid=9 rel=1 In 1969, the United States became the first nation to land men on the surface of the Moon.
10 consine=0.6250 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
11 consine=0.1581 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County.
12 consine=0.3627 rank=3 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
13 consine=0.2236 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 consine=0.5270 rank=1 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 consine=0.1788 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.
16 consine=0.2938 rank=3 qid=16 rel=1 Bob Marley died in 1981 at age 36.
17 consine=0.2041 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which from a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers.
18 consine=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.
19 consine=0.1650 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.
20 consine=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.
21 (MMR) Mean Reciprocal Rank ::0.5375

```

Figure 2. Result of new tokenization algorithm

## 2.2 Bad stemming algorithm

The simple vector model I used in task 1 cannot know that different tense of a same verb should mean the same thing. Here are examples in impute data that simple vector model do not perform well.

qid	rel	error	correct
8	99 and 1	'bite' not match 'bit'	'bite' and 'bit' have same meaning
16	99 and 1	'die' not match 'died'	'die' and 'died' have same meaning

Table 2.

In this case, simple vector model should be improved. I use Stanford lemmatizer to improve stemming algorithm. Figure 3 shows the result of this improved algorithm. After using this algorithm, the MMR became 0.55 from 0.4375(compared with original result, not after using new tokenization algorithm).

```

1 consine=0.2667 qid=2 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. Michael Jordan, one of the greatest basketball players of all time - made what was expected to be his last trip to play basketball from August in year 1999.
2 consine=0.2663 qid=1 rel=1 Michael Jordan, one of the greatest basketball players of all time - made what was expected to be his last trip to play basketball from August in year 1999.
3 consine=0.4714 qid=3 rel=1 Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
4 consine=0.3086 qid=1 rel=4 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
5 consine=0.0999 qid=3 rel=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China".
6 consine=0.5547 qid=2 rel=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7 consine=0.0891 qid=4 rel=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
8 consine=0.2750 qid=2 rel=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
9 consine=0.2754 qid=2 rel=9 rel=1 The first man to walk on the surface of the Moon.
10 consine=0.7500 qid=10 rel=10 rel=1 Menchu won the Nobel peace prize in 1992.
11 consine=0.3536 qid=2 rel=11 rel=1 Devils Tower can be found in Crook County.
12 consine=0.3162 qid=4 rel=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
13 consine=0.1195 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 consine=0.4216 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 consine=0.7727 qid=3 rel=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded in East Antarctica.
16 consine=0.4938 qid=16 rel=1 Bob Marley died in 1981 at age 36.
17 consine=0.3015 qid=3 rel=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers .
18 consine=0.2265 qid=2 rel=18 rel=1
19 consine=0.2417 qid=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
20 consine=0.3078 qid=2 rel=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
21 (MR) Mean Reciprocal Rank ::0.55

```

Figure 3. Result of new stemming algorithm

## 2.3 simple similarity measure

In task 1, the similarity measure I used is cosine similarity, which is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. If cosine similarity is expressed over bit vectors  $A(a_1, a_2, \dots, a_n)$  and  $B(b_1, b_2, \dots, b_n)$ , then it can be written as:

$$F(A, B) = \frac{A \cdot B}{\sqrt{|A|^2} \times \sqrt{|B|^2}} = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \times \sqrt{\sum_{i=1}^n b_i^2}}$$

However, I'm not sure that whether this method is good enough or not. So I implement other two similarity measure methods: Jaccard coefficient and Dice coefficient.

### 2.3.1 Jaccard coefficient

The Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. If Jaccard similarity is expressed over bit vectors  $A(a_1, a_2, \dots, a_n)$  and  $B(b_1, b_2, \dots, b_n)$ , then it can be written as:

$$F(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} = \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n a_i^2 + \sum_{i=1}^n b_i^2 - \sum_{i=1}^n a_i b_i}$$

Figure 4 shows the result of using Jaccard instead of cosine similarity. After using it, the MMR became 0.4625 from 0.4375(implemented on task 1 and compared with original result). The MMR did not improve rapidly, but it still shows better result.

```

Applications Places Mon Oct 20, 5:37 PM
Java - hw3-ningnaw/report.txt - Eclipse
File Edit Navigate Search Project Run Window Help
File Navigator Search Project Run Window Help
report.txt documents.txt RetrievalEvaluator.java DocumentVectorAnnotator.java
1.Jaccard=0.1622 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.
2.Jaccard=0.1273 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.
3.Jaccard=0.1333 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
4.Jaccard=0.1333 rank=2 qid=4 rel=1 On March 2, 1972, Wilt Chamberlain scored record 100 points in a game against the New York Knicks.
5.Jaccard=0.0000 rank=1 qid=5 rel=1 Some of China have had foghogs about River, which they often call "sorrow of China".
6.Jaccard=0.3810 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7.Jaccard=0.0417 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.
8.Jaccard=0.0909 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.
9.Jaccard=0.3810 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
10.Jaccard=0.3333 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1992.
11.Jaccard=0.0000 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County.
12.Jaccard=0.1618 rank=12 qid=12 rel=1 The world's oldest, 4,000 year old redwood stands 367 1/2 feet tall.
13.Jaccard=0.0556 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14.Jaccard=0.2667 rank=1 qid=14 rel=1 Lionel Ritchie was lead singer and songwriter for Commodores.
15.Jaccard=0.0345 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.
16.Jaccard=0.1538 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17.Jaccard=0.0714 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which is the nation's top producer.
18.Jaccard=0.0588 rank=3 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers used to run their business spread across the country.
19.Jaccard=0.0588 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burst into flames and crashed in Lakehurst, N.J., killing 35 of the 96 passengers and crew.
20.Jaccard=0.1429 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.
21. (MMR) Mean Reciprocal Rank ::0.4625

```

Figure 4. Result of Jaccard coefficient

### 2.3.2 Dice coefficient

Dice coefficient is also a statistic used for comparing the similarity of two samples. If Dice coefficient is expressed over bit vectors  $A(a_1, a_2, \dots, a_n)$  and  $B(b_1, b_2, \dots, b_n)$ , then it can be written as:

$$F(A, B) = \frac{2(A \cdot B)}{|A|^2 + |B|^2} = \frac{2 \sum_{i=1}^n a_i b_i}{\sum_{i=1}^n a_i^2 + \sum_{i=1}^n b_i^2}$$

Figure 5 shows the result of using Dice instead of cosine similarity. After using it, the MMR became 0.4625 from 0.4375(implemented on task 1 and compared with original result), which has same result with implementing Jaccard coefficient. Maybe it is because the sample input is typical in these two implementations.

The screenshot shows the Eclipse IDE interface with a text editor open containing a list of 21 entries. Each entry consists of a Dice coefficient value followed by a rank, a query ID (qid), a relation ID (rel), and a descriptive sentence. The sentences cover various historical and scientific events.

```

1Dice=0.2791 rank=2 qid=1 rel=l In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an estimated 20,000 people died.
2Dice=0.2363 rank=3 qid=2 rel=l Michael Jordan, one of the greatest basketball player of all time--made what was expected to be his last trip to play.
3Dice=0.2353 rank=3 qid=3 rel=l Alaska was purchased from Russia 1 year ago.
4Dice=0.2069 rank=2 qid=4 rel=l On March 2, 1962, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
5Dice=0.0000 rank=3 qid=5 rel=l People of China have mixed feelings about River, which they often call "sorrow of China".
6Dice=0.5517 rank=2 qid=6 rel=l Roger Bannister was the first to break the four-minute mile barrier.
7Dice=0.0800 rank=3 qid=7 rel=l And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
8Dice=0.1667 rank=4 qid=8 rel=l Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
9Dice=0.0000 rank=5 qid=9 rel=l The first man to walk on the Moon was Neil Armstrong.
10Dice=0.5000 rank=1 qid=10 rel=l Menchu won the Nobel peace prize in 1992.
11Dice=0.1667 rank=4 qid=11 rel=l Devils Tower can be found in Crook County.
12Dice=0.3077 rank=3 qid=12 rel=l Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
13Dice=0.1053 rank=3 qid=13 rel=l Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14Dice=0.4211 rank=4 qid=14 rel=l Lionel Ritchie was lead singer and songwriter for Commodores.
15Dice=0.0689 rank=3 qid=15 rel=l A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded in East Antarctica.
16Dice=0.0667 rank=3 qid=16 rel=l Bed Mats are distributed in Japan.
17Dice=0.1333 rank=3 qid=17 rel=l Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which often has dry springs.
18Dice=0.1379 rank=2 qid=18 rel=l From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the McDonald brothers developed spread across the nation.
19Dice=0.1111 rank=3 qid=19 rel=l On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakehurst, N.J., killing 35 of the 91 passengers and crew.
20Dice=0.2500 rank=2 qid=20 rel=l They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
21 (MMR) Mean Reciprocal Rank ::0.4625

```

Figure 5. Result of Dice coefficient

## 2.4 Total System Improvement

Based on all individual improvement I did from 2.1 to 2.3, I came up with an idea that I can combine these improvements together. So I implement it. And result shows that this idea really performed well. It improved MMR from 0.4375 to 0.6833. And Figure 6 shows UML diagram of my implementation as well as Figure 7 shows result of this implementation.

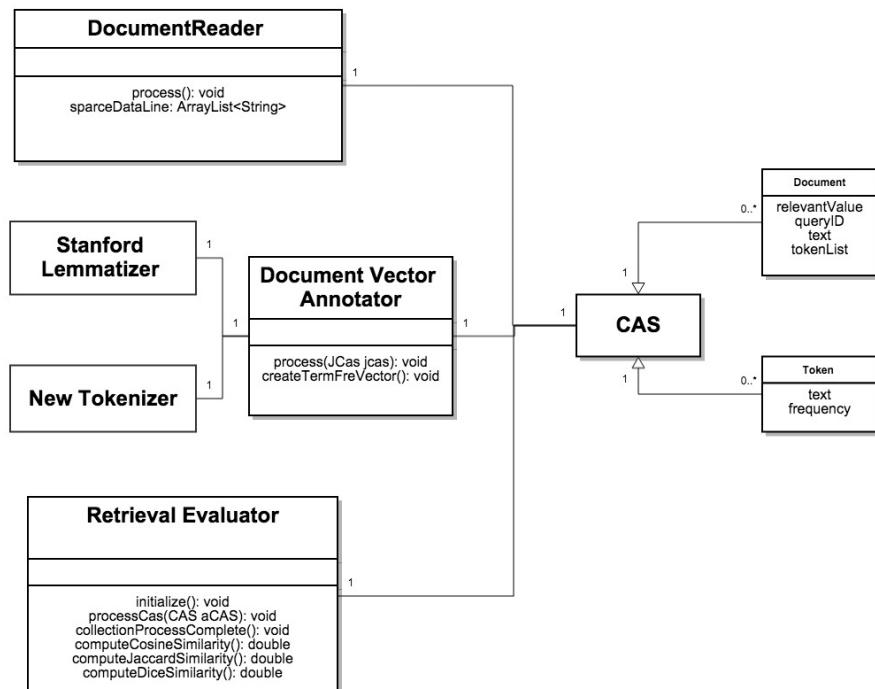


Figure 6. UML Diagram

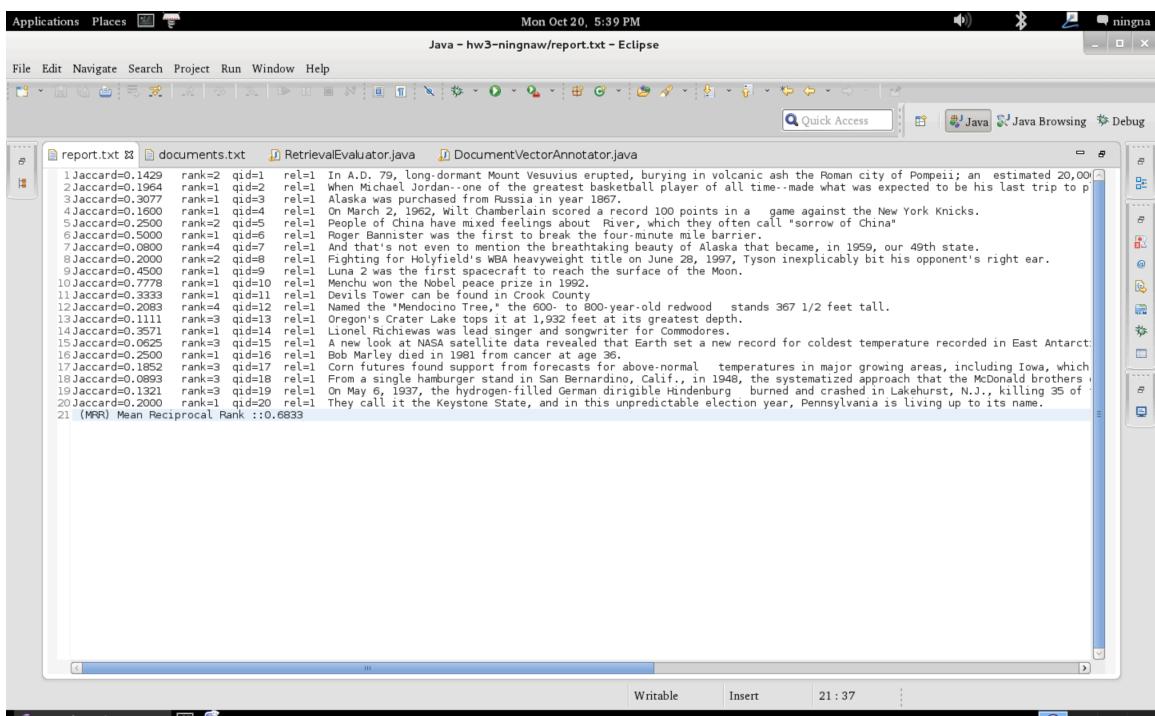


Figure 7. Result of Total Improvement