

Andrew id: qiqis  
October 18, 2014

# hw3-report

## Error Analysis

The following picture is the result by using default tokenization algorithm, cosine similarity method without stemming algorithms. MRR is 0.4375 according to these algorithms.

```
1 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; on
2 cosine=0.2818 rank=7 qid=7 rel=1 When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be hi
3 cosine=0.2357 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 cosine=0.7315 rank=7 qid=4 rel=1 On March 7, 1967, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
5 cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
6 cosine=0.5547 rank=7 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7 cosine=0.0091 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 cosine=0.1813 rank=7 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably hit his opponent's
9 cosine=0.5804 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
10 cosine=0.5000 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1997.
11 cosine=0.1768 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County
12 cosine=0.3167 rank=1 qid=12 rel=1 Named the "Mendocino Tree," the 680- to 800-year-old redwood stands 367 1/2 feet tall.
13 cosine=0.1195 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 cosine=0.4716 rank=7 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 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 records
16 cosine=0.2878 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.1988 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, incl
18 cosine=0.7765 rank=7 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the M
19 cosine=0.1268 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakshurst, N.J
20 cosine=0.3878 rank=7 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to
21 MRR=0.4375
```

## Stemming Mismatch

As the figure shows below, I implement another way of tokenization by using Stanford Lemmatizer. By using this method, MRR is increased to 0.5500 which means that due to the Stemming ignorance, the original method does not performance well. Therefore, we can say that Stemming mismatch is a error type.

```
1 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; on
2 cosine=0.0091 rank=1 qid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be hi
3 cosine=0.4714 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 cosine=0.3886 rank=1 qid=4 rel=1 On March 7, 1967, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
5 cosine=0.0990 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
6 cosine=0.5547 rank=7 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7 cosine=0.0091 rank=7 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
8 cosine=0.7750 rank=7 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably hit his opponent's
9 cosine=0.5804 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
10 cosine=0.7500 rank=1 qid=10 rel=1 Menchu won the Nobel peace prize in 1997.
11 cosine=0.3536 rank=2 qid=11 rel=1 Devils Tower can be found in Crook County
12 cosine=0.3167 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 680- to 800-year-old redwood stands 367 1/2 feet tall.
13 cosine=0.1195 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 cosine=0.4716 rank=3 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 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 records
16 cosine=0.4743 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.3015 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, incl
18 cosine=0.7765 rank=7 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the M
19 cosine=0.2417 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakshurst, N.J
20 cosine=0.3878 rank=7 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to
21 MRR=0.5500
```

Take the qid=3 for example. The original rank is 3. After the Stemming match, the rank is 1. The original one does not match word “purchase” with “purchased” which affects its performance.

qid=3 rel=99 In which year did a purchase of Alaska happen?  
qid=3 rel=1 Alaska was purchased from Russia in year 1867.

Therefore, for 5 queries, the most relevant answers were not selected due to Stemming Mismatch.

## Tokenization Mistake

As the figure shows below, I implement another way of tokenization by using Stanford Tokenizer. By using this method, MRR is increased to 0.5125 which means that due to the Tokenization mistakes, the original method does not performance well. Therefore, we can say that tokenization mistake is a error type.

```
cosine=0.2872 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
cosine=0.3837 rank=1 qid=7 rel=1 When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be hi
cosine=0.2188 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.7615 rank=1 qid=4 rel=1 On March 7, 1967, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
cosine=0.1588 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
cosine=0.5345 rank=7 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.1387 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=0.3769 rank=7 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably hit his opponent's
cosine=0.5657 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5556 rank=1 qid=10 rel=1 Manchu won the Nobel peace prize in 1997.
cosine=0.1581 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.3371 rank=3 qid=12 rel=1 Named the "Mendocino Tree," the 680- to 800-year-old redwood stands 367 1/2 feet tall.
cosine=0.2041 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
cosine=0.5888 rank=1 qid=14 rel=1 Lionel Ritchie was lead singer and songwriter for Commodores.
cosine=0.8722 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded
cosine=0.7697 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.1898 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, inclu
cosine=0.7745 rank=7 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the M
cosine=0.1368 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakehurst, N.J
cosine=0.3333 rank=7 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to
MRR=0.5125
```

Take the qid=2 for example. The original rank is 2. After the Stanford Tokenization algorithm, the rank is 1. The original one does not split word according the punctuation such as “-” which ignores the similarity between Jordan and affects its performance.

qid=2 rel=99 What has been the largest crowd to ever come see  
Michael Jordan  
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.

Therefore, for 3 queries, the most relevant answers were not selected due to Tokenization Mistake.

## Similarity Measure Flaw

### Jaccard Similarity

As the figure shows below, I implement another way of calculating similarity by using Jaccard Similarity way. By using this method, MRR is increased to 0.4875 which means that due to the choice of different similarity calculation way, the original method does not performance well. Therefore, we can say that Similarity Measure Flaw is a error type.

1 cosine=0.1826	rank=2	qid=1	rel=1	In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; on
2 cosine=0.0893	rank=7	qid=2	rel=1	When Michael Jordan--one of the greatest basketball player of all time--made what was expected to be his
3 cosine=0.1853	rank=3	qid=3	rel=1	Alaska was purchased from Russia in year 1867.
4 cosine=0.1798	rank=7	qid=4	rel=1	On March 7, 1967, Walt Chamberlain scored a record 100 points in a game against the New York Knicks.
5 cosine=0.0385	rank=3	qid=5	rel=1	People of China have mixed feelings about River, which they often call "sorrow of China"
6 cosine=0.2388	rank=2	qid=6	rel=1	Roger Bannister was the first to break the four-minute mile barrier.
7 cosine=0.0667	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 cosine=0.1798	rank=7	qid=8	rel=1	Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably hit his opponent's
9 cosine=0.1818	rank=2	qid=9	rel=1	Luna 2 was the first spacecraft to reach the surface of the Moon.
10 cosine=0.2778	rank=1	qid=10	rel=1	Menchu won the Nobel peace prize in 1997.
11 cosine=0.0769	rank=4	qid=11	rel=1	Devils Tower can be found in Crook County
12 cosine=0.1111	rank=1	qid=12	rel=1	Named the "Mendocino Tree," the 480- to 800-year-old redwood stands 367 1/2 feet tall.
13 cosine=0.1800	rank=3	qid=13	rel=1	Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 cosine=0.2778	rank=1	qid=14	rel=1	Lionel Richie was lead singer and songwriter for Commodores.
15 cosine=0.0323	rank=3	qid=15	rel=1	A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded
16 cosine=0.1758	rank=1	qid=16	rel=1	Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.0857	rank=3	qid=17	rel=1	Corn futures found support from forecasts for above-normal temperatures in major growing areas, inclu
18 cosine=0.0417	rank=7	qid=18	rel=1	From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the M
19 cosine=0.0625	rank=3	qid=19	rel=1	On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakehurst, N.J
20 cosine=0.4875	rank=1	qid=20	rel=1	They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to
21 MRR=0.4875				

Take the qid=20 for example. The original rank is 2. After the Jaccard Similarity, the rank is 1.

qid=20 rel=99 What is the Keystone State?  
qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.

Therefore, for 2 queries, the most relevant answers were not selected due to Similarity Measure Flaw.

### Dice Coefficient Similarity

As the figure shows below, I implement another way of calculating similarity by using Dice coefficient Similarity method. By using this method, MRR is increased to 0.4875 which means that due to the choice of different similarity calculation method, the



original method does not performance well. Therefore, we can say that Similarity Measure Flaw is a error type.

```
1 cosine=0.7851 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
2 cosine=0.1786 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
3 cosine=0.7185 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 cosine=0.2581 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.
5 cosine=0.8709 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"
6 cosine=0.4615 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
7 cosine=0.1333 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 cosine=0.2581 rank=2 qid=8 rel=1 Fighting for Molyfield's NBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's
9 cosine=0.3636 rank=2 qid=9 rel=1 Luna 7 was the first spacecraft to reach the surface of the Moon.
10 cosine=0.5556 rank=1 qid=10 rel=1 Monchu won the Nobel peace prize in 1992.
11 cosine=0.1538 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County
12 cosine=0.2222 rank=3 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
13 cosine=0.7888 rank=3 qid=13 rel=1 Oregon's Crater Lake tops it at 1,937 feet at its greatest depth.
14 cosine=0.5556 rank=1 qid=14 rel=1 Lionel Richie was lead singer and songwriter for Commodores.
15 cosine=0.8567 rank=3 qid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded
16 cosine=0.2500 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.1714 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatures in major growing areas, inclu
18 cosine=0.8833 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized approach that the M
19 cosine=0.1758 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and crashed in Lakehurst, N.J
20 cosine=0.2857 rank=1 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to
MRR=0.4875
```

Take the qid=20 for example. The original rank is 2. After the Dice coefficient Similarity, the rank is 1.

qid=20 rel=99 What is the Keystone State?  
qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.

Therefore, for 2 queries, the most relevant answers were not selected due to Similarity Measure Flaw.

## Others

For the ties which means that there are some same similarity which cases ambiguous in rank. In this way, the MRR is affected by the error.

At the same time, different token sequence can express different meanings. However, the cosine similarity does not take this into consideration. So in this way, this can affect the rank of different tokens.

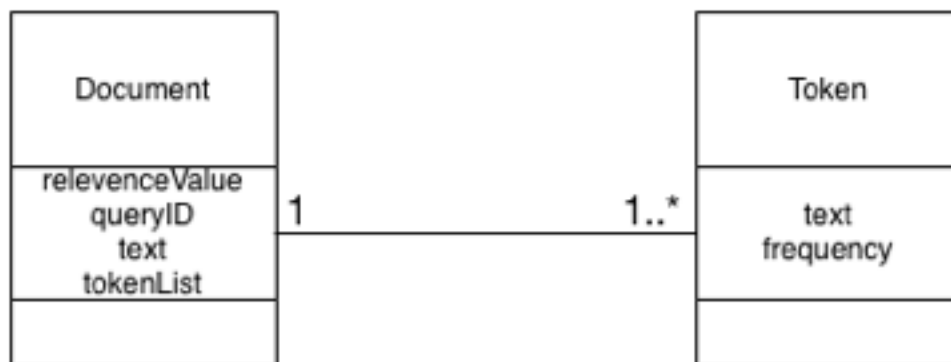
## Conclusion

	Error Type	Frequency
1	Stemming Mismatch	5
2	Tokenization Mistake	3
3	Similarity Measure Flaw	2
4	Ties	3
5	Token Sequence	10

## Architecture

### Type system

The following figure shows the data structure of this program.



Document is the structure to store the relevanceValue, queryID, text and tokenList for each document.

Token is to store the token with the relative frequency for each document.

In this way, one Document is corresponding to one or more Tokens.

### Data structure

For TokenVector: I use HASHMAP to store the token vector for each document. And because of high efficiency of hash map. find function and this will highly improve the performance of the program.

```
/** store the term and term frequency */  
HashMap<String, Integer> hs= new HashMap<String, Integer>();  
|
```

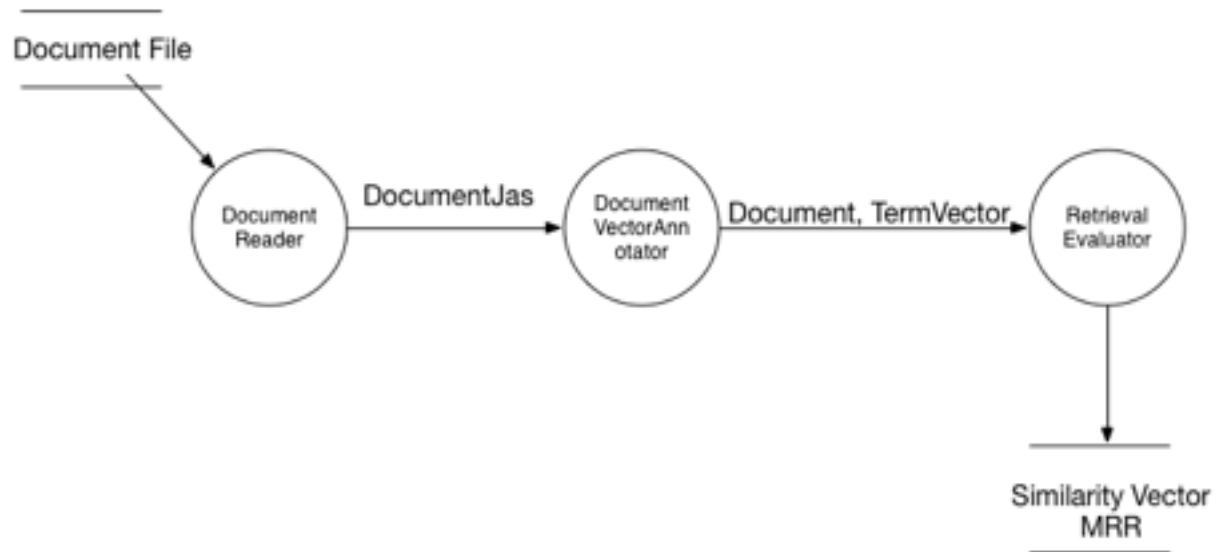
For similarity for each document and group, I used the following structure to store.

```
/** similarity for all */  
public ArrayList<ArrayList<Double>> similarityAll=new ArrayList<ArrayList<Double>>();
```

For document information, I use a class DocInfo to store all the related info about Document as the figure shows below:

```
class DocInfo  
{  
    int qid;  
    int rel;  
    Map<String, Integer> vector;  
    String text;  
  
    public DocInfo(int qid,int rel,Map<String,Integer> vector,String text) {  
        // 1000 Auto-generated constructor stub  
        this.qid=qid; //the id for the documents  
        this.rel=rel; // the relevant for each documents  
        this.vector=vector; //the vector of the tokenlist and frequency  
        this.text=text; //the text for each documents  
    }  
  
    public Map<String, Integer> GetVector()  
    {  
        return vector;  
    }  
  
    public int GetQid()  
    {  
        return qid;  
    }  
  
    public int GetRel()  
    {  
        return rel;  
    }  
  
    public String GetText()  
    {  
        return text;  
    }  
}
```

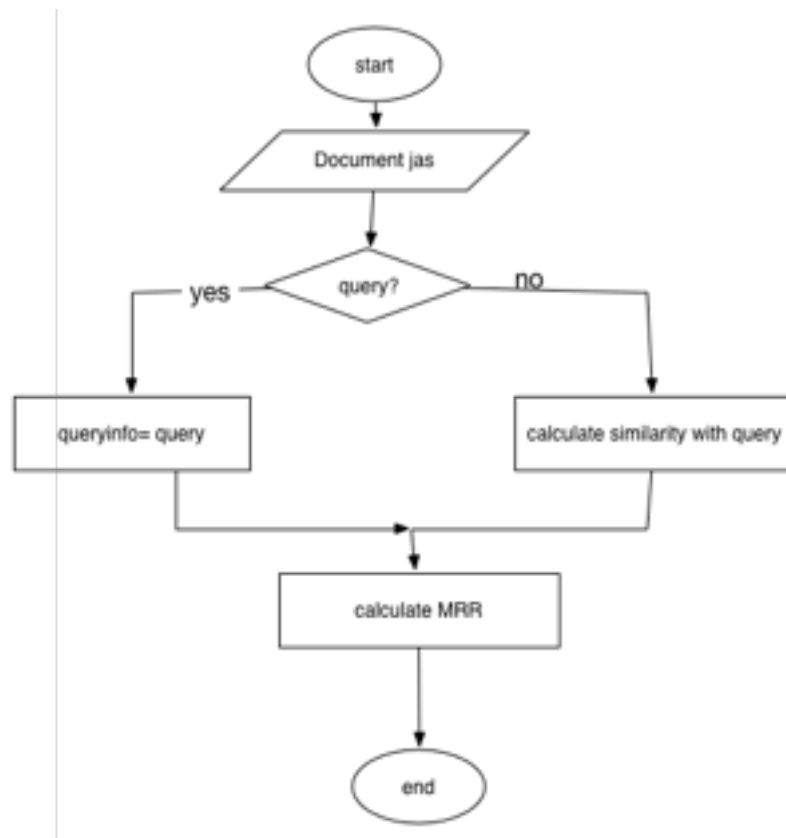
## System structure



The system data flow is shown as the figure above. And DocumentReader is responsible for reading document data from file and store in DocumentJas. DocumentVectorAnnotator is responsible for analyzing the term frequency of each term and create Term vector for each document. RetrievalEvaluator is responsible for calculate the similarity between each document and give the rank, finally calculating the MRR.

## Flow for calculating the similarity

Without using a complex structure to store all data from file, I calculated the similarity between each query and document during the traversal.



## Knowledge Base

### Tokenization algorithm- Stanford Tokenization

PTBTokenizer is an efficient, fast, deterministic tokenizer. On a typical 2010 computer, it will tokenize text at a rate of about 200,000 tokens per second.

This is much more efficient than the tokenization function which only considers the whitespace.

### Stemming algorithm- StanfordLemmatizer

The simple vector model does not understand that 'die' and 'dies' denote the same thing in most contexts. StanfordLemmatizer is used to stem words more efficiently.



## **Similarity Algorithm**

### **Jaccard Index**

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

The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets

### **Dice coefficient**

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

A and B are the number of species in samples A and B, respectively, and C is the number of species shared by the two samples; QS is the quotient of similarity and ranges from 0 to 1.