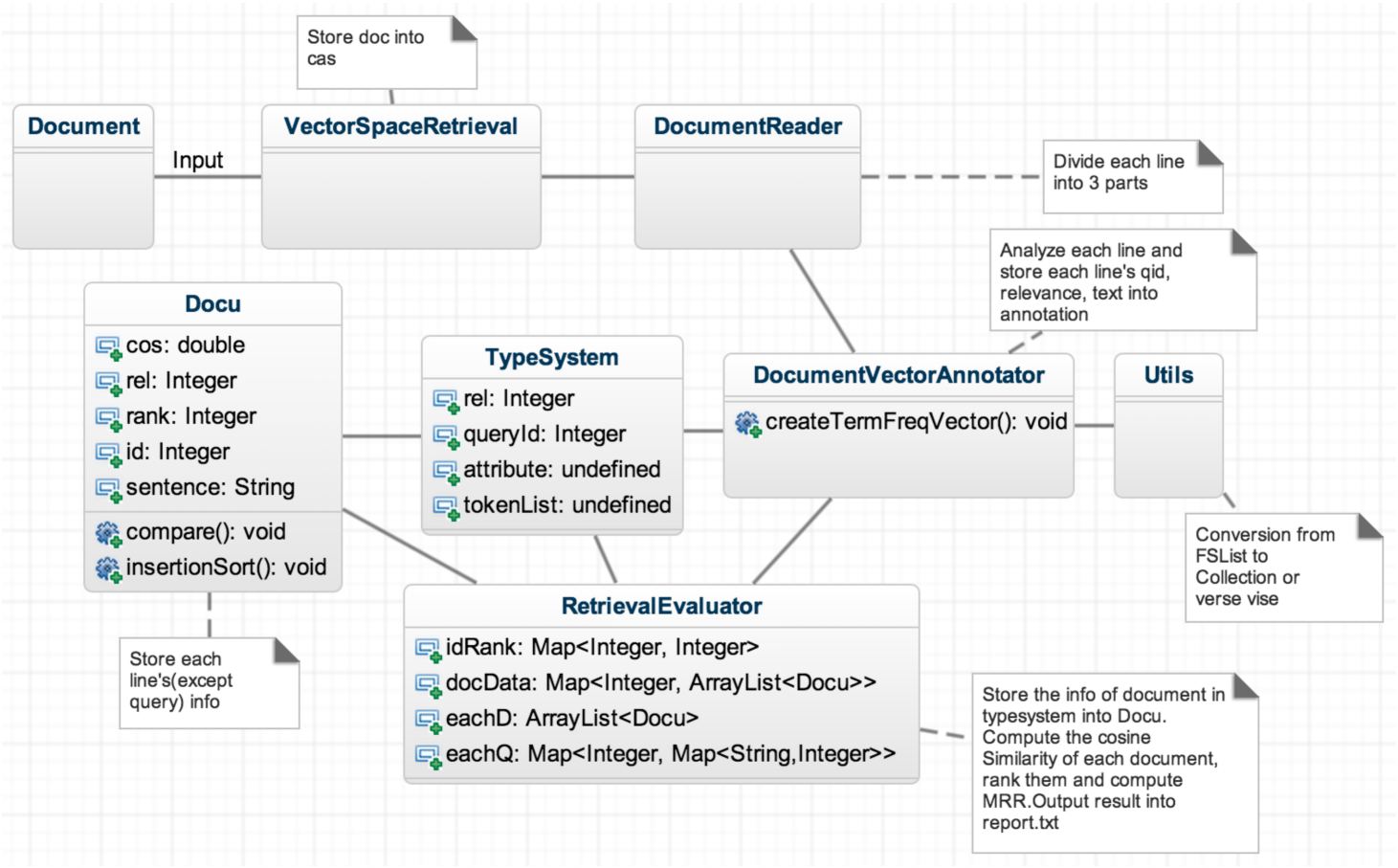
# Task 1 Design

The main structure is demonstrated in the model diagram below.




## Docu

I build a new Class “Docu” to store the info of each line.
It also contain two function “compare()” which is used to compare two Docu instance by their cosine.
insertionSort(), which is used to sort a list of array containing Docu.

## RetrievalEvaluator

In this class, I build several Map to store info of query and document in memory.
In addition, implement the function of calculation of mmr and cosine similarity.

# Performance

```
cosine=0.2791   rank= 2   qid= 1   rel=1   In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcar
cosine=0.2858   rank= 2   qid= 2   rel=1   When Michael Jordan--one of the greatest basketball player of all
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 c
cosine=0.0000   rank= 3   qid= 5   rel=1   People of China have mixed feelings about  River, which they ofter
cosine=0.5547   rank= 2   qid= 6   rel=1   Roger Bannister was the first to break the four-minute mile barrie
cosine=0.0891   rank= 3   qid= 7   rel=1   And that's not even to mention the breathtaking beauty of Alaska t
cosine=0.1833   rank= 2   qid= 8   rel=1   Fighting for Holyfield's WBA heavyweight title on June 28, 1997, T
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   st
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 r
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   temp
cosine=0.2265   rank= 2   qid= 18   rel=1   From a single hamburger stand in San Bernardino, Calif., in 1948,
cosine=0.1268   rank= 3   qid= 19   rel=1   On May 6, 1937, the hydrogen-filled German dirigible Hindenburg
cosine=0.3078   rank= 2   qid= 20   rel=1   They call it the Keystone State, and in this unpredictable electi
 (MRR) Mean Reciprocal Rank ::0.4374999999999999
```

# Task 2

## 2. 1 Error analysis

Since Cosine similarity only concerns whether words in document

## 2.1.1 Vocabulary mismatch

### 2.1.1.1 Singular & plural

For example, cities and city.

### 2.1.1.2 Upper & lower

Words that are the start of a sentence may have this problem.

### 2.1.1.3 Tense

Different tense or verb. For example, “get” and “got”.

#### 2.1.1.4 Abbreviate

Abbreviation of words. For example “I’ve” and “I have”.

#### 2.1.1.5 Abbreviate

Punctuation differences. Since we only divide word by space, the punctuation is included in the word before.

## 2.1.2 Irrelevant match words with query in incorrect answers

### 2.1.2.1 Stopwords

Some words may be irrelevant to the key words, have little contribute to the main topic. Too much such words in incorrect answer may increase the weight of a document.
For example, “of, are”

### 2.1.2.2 Same words, different topics

The words in wrong answer may be the same keywords in the query, thus making this document’s weight very high.
In fact, every wrong answer have this factor.

## 2.1.3 Irrelevant mismatch

### 2.1.3.1 Too much detail

There are to many irrelevant information in the correct answer, thus making it’s score very low.
For example, in Query 17:
**Q:** Which U.S. state is the leading corn producer?
**A:** Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which often is the nation's largest producer.
The green part is the irrelevant details, most of which don’t exist in the query.

## 2.1.4 Synonyms

The right answer does tell the right thing, but it may tell in the different way, using different words from ones in query.

# 2.2 Sheet

There are three major kinds of factors influencing the performance of our system.
They are as below.

| Error type | Specific class | Example | Frequency |
|---|---|---|---|
| Vocabulary mismatch | 1. Singular & plural,<br>2. Upper & lower.<br>3. Tense<br>4. Abbreviate.<br>5. Punctuation<br>6. Typo | 1. q1, q4,<br>2. q2, q17,<br>3. q3, q4, q8, q11, q15, q16<br>4. q11, q18,<br>5. q1, q2, q6, q9, q12, q14, q15, q17<br>6. | 1. 2<br>2. 2<br>3. 6<br>4. 2<br>5. 8<br>6. 0 |
| Irrelevant match words in incorrect answers | 1. Of, the, are, etc.<br>2. Same words, different topics | 1. q6, q7, q13, q14, q15, q17, q18, q20<br>2. All | 1. 8<br>2. 20 |
| Irrelevant mismatch | 1. Too much detail | 1. q1, q2, q4, q5, q7, q8, q12, q13, q14, q15, q16, q17, q18, q19, q20. | 1. 14 |
| Synonyms | | q1, q6, q9, q11, q12, q13, q15, q17, q18, q19, q20. | 11 |

Q: query, A: answer, WA: wrong answer

# Query 1

**Q:** Give us the name of the volcano that destroyed the ancient city of Pompeii
**A:** In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Roman city of Pompeii; an estimated 20,000 people died.
**WA:** Vesuvius is located near the ruins of the destroyed city of Pompeii.

# Query 2

**Q:** What has been the largest crowd to ever come see Michael Jordan
**A:** 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.
**WA:** A supposedly last play of Michael Jordan gathered some of the largest crowd in history of NBA.

# Query 3

**Q:** In which year did a purchase of Alaska happen?
**A:** Alaska was purchased from Russia in year 1867.
**WA:** William Seward negotiated a purchase of Alaska for $7.2 million.

# Query 4

**Q:** What year did Wilt Chamberlain score 100 points?
**A:** On March 2, 1962, Wilt Chamberlain scored a record 100 points in a game against the New York Knicks.
**WA:** A 100 point game was a highlight in a career of Wilt Chamberlain

# Query 5

**Q:** What river is called China's Sorrow?
**A:** People of China have mixed feelings about River, which they often call "sorrow of China"
**WA:** Yellow river is often called the mother of China

# Query 6

**Q:** Who was the first person to run the mile in less than four minutes
**A:** Roger Bannister was the first to break the four-minute mile barrier.
**WA:** It is hard for humans to run the mile faster than in four minutes

# Query 7

**Q:** What year did Alaska become a state?
**A:** And that's not even to mention the breathtaking beauty of Alaska that became, in 1959, our 49th state.
**WA:** Also as it did in 1959, the state of Alaska was struggling with how to pay its bills.

# Query 8

**Q:** When did Mike Tyson bite Holyfield's ear?
**A:** Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his opponent's right ear.
**WA:** Tyson was choked up as he spoke of biting Evander Holyfield's ear.

# Query 9

**Q:** What was the first spaceship on the moon
**A:** Luna 2 was the first spacecraft to reach the surface of the Moon.
**WA:** Eagle was the first manned spacecraft that reached the surface of the moon

# Query 11

**Q:** Where is Devil's Tower
**A:** Devils Tower can be found in Crook County
**WA:** Devil's Tower is an igneous intrusion that rises dramatically 1,267 feet (386 m) above the surrounding terrain.

# Query 12

**Q:** What is the height of the tallest redwood?
**A:** Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
**WA:** Mendocino Tree is the tallest redwood in the world.

# Query 13

**Q:** How deep is Crater Lake?
**A:** Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
**WA:** Crater Lake is a caldera lake in the western United States.

# Query 14

**Q:** Who was the lead singer for the Commodores
**A:** Lionel Richiewas was lead singer and songwriter for Commodores.
**WA:** The Commodores originally came together from groups the Mystics and the Jays.

# Query 15

**Q:** What is the coldest place on earth?
**A:** A new look at NASA satellite data revealed that Earth set a new record for coldest temperature recorded in East Antarctica.
**WA:** Oymyakon is the coldest place in Russia.

# Query 16

**Q:** When did Bob Marley die
**A:** Bob Marley died in 1981 from cancer at age 36.
**WA:** Bob Marley was a Jamaican reggae singer-songwriter, musician, and guitarist who did achieve international fame.

# Query 17

**Q:** Which U.S. state is the leading corn producer?
**A:** Corn futures found support from forecasts for above-normal temperatures in major growing areas, including Iowa, which often is the nation's largest producer.
**WA:** The United States is the world's leading producer of corn

# Query 18

**Q:** Where was the first McDonald's built?
**A:** 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.
**WA:** McDonald's Corporation is the world's largest chain of hamburger fast food restaurants.

# Query 19

**Q:** The Hindenburg disaster took place in 1937 in which New Jersey town?
**A:** 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.
**WA:** The Hindenburg disaster took place as the German passenger airship LZ 129 Hindenburg caught fire and was destroyed during its attempt to dock with its mooring mast

# Query 20

**Q:** What is the Keystone State?

**A:** They call it the Keystone State, and in this unpredictable election year, Pennsylvania is living up to its name.
**WA:** Keystone Resort is the largest ski resort in Summit County located in Keystone Colorado.

Since q10 gets the right answer, I don't include it.

# 2.3 Improvement methods

## 2.3.1 Better Tokenization algorithm

Remove punctuations.

## 2.3.2 Better stemming algorithm:

Translate all the word into lower case and the same tense.
Remove plural.
Remove abbreviation.

## 2.3.3 Better or different similarity measures:

Exclude stop words.
Use BM25 to judge. Since cosine similarity cannot tell "John loves Mary" and "Mary loves John" which one is right

# 2.4 Improvement design and Implementation

## 2.4.1 BM25

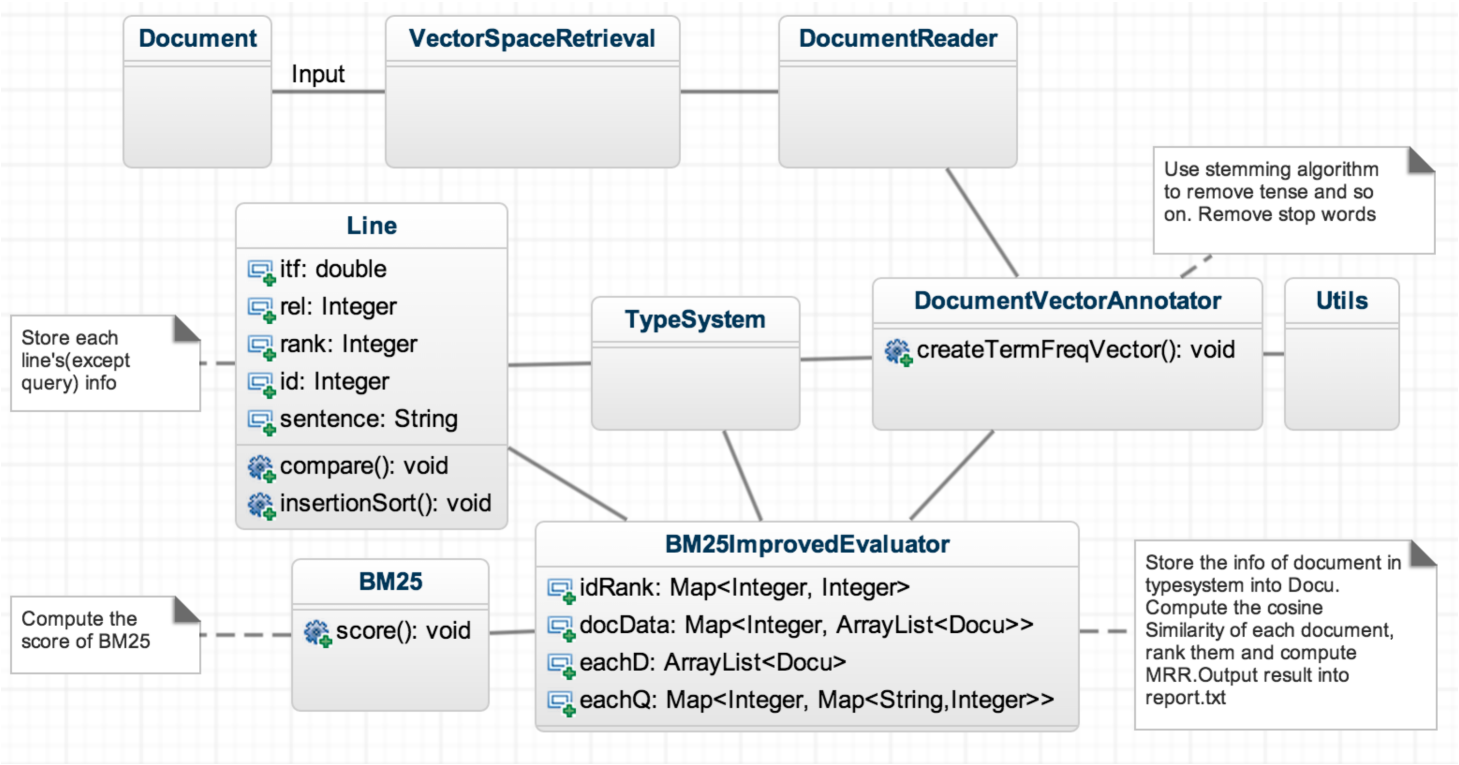
BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.
Given a query "Q", containing keywords "q1, q2, q3… qn", the BM25 score of a document "D" is:

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

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

## Design

The design is as follow.




I only describe modifications compared with the original version

## Line

I build another Class to store information of document. It is similar to the class “Docu”, just with different variable “itf”

## DocumentVectorAnnotator

In this class, when storing info from cas to typesystem, I used function to remove all the punctuation, and also use StanfordLemmatizer to remove all the stemming problems.

## BM25

I use this class to compute the itf score using the function score().

## BM25ImprovedEvaluator

In this class, I use similar data structure to store info of doc and query in memory. Just have to iterate through query's word to compute its term frequency and several variables demanded to compute BM25.

## Performance

```
itf=-3.0769     rank=1  qid=13  Oregon's crater Lake tops it at 1,932 feet at its greatest depth.
itf=-0.9243     rank=1  qid=14  Lionel Richiewas was lead singer and songwriter for Commodores.
itf=-1.0489     rank=1  qid=15  A new look at NASA satellite data revealed that Earth set a new record for cc
itf=-5.0186     rank=3  qid=16  Bob Marley died in 1981 from cancer at age 36.
itf=-2.1619     rank=2  qid=17  Corn futures found support from forecasts for above-normal   temperatures in
itf=-1.4407     rank=1  qid=18  From a single hamburger stand in San Bernardino, Calif., in 1948, the systemc
itf=-2.2913     rank=2  qid=19  On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and
itf=0.0000      rank=2  qid=20  They call it the Keystone State, and in this unpredictable election year, Per
 (MRR) Mean Reciprocal Rank ::0.5625
Total time taken: 0.997
```

Score 0.4375 -> 0.5625

## Advantage

The performance does increase. There is a parameter b in this algorithm, which could be adjusted due to different length of document. So the algorithm is flexible.
When one word shows up in several answers, means incorrect answer may contain the same keyword, which could be very interferential in cosine similarity. But BM25 will decrease the weight of such words, to reduce disturbs of such incorrect answers.

## Disadvantage

It lower the weight of keyword, in some cases may lead to wrong answer, since the correct word may have many keywords

# 2.4.1 Jaccard coefficient method

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. 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:

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

By using this method, we may not get a better Mean Reciprocal Rank (MRR) than the cosine similarity. It just simply gets the intersection set and the union set.

## Performance

```
jaccard=0.0007  rank=1  qid=13  Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
jaccard=0.3333  rank=1  qid=14  Lionel Richiewas was lead singer and songwriter for Commodores.
jaccard=0.0385  rank=1  qid=15  A new look at NASA satellite data revealed that Earth set a new record for cc
jaccard=0.1538  rank=1  qid=16  Bob Marley died in 1981 from cancer at age 36.
jaccard=0.0690  rank=1  qid=17  Corn futures found support from forecasts for above-normal   temperatures in
jaccard=0.0278  rank=1  qid=18  From a single hamburger stand in San Bernardino, Calif., in 1948, the systemc
jaccard=0.0556  rank=1  qid=19  On May 6, 1937, the hydrogen-filled German dirigible Hindenburg   burned and
jaccard=0.1429  rank=1  qid=20  They call it the Keystone State, and in this unpredictable election year, Per
 (MRR) Mean Reciprocal Rank ::0.4417
Total time taken: 0.958
```

Score 0.4375 -> 0.4417

## Design

The design is similar to cosine similarity, which is shown in the task 1. Just replace the function of cosine similarity with jaccard.

## Advantage

The performance does increase, but not that evident.
It doesn't need normalized.
Since it is similar to cosine similarity, the improvement may due to erase the punctuation and stop words.

## Disadvantage

Since it is similar to cosine similarity, it still faces problems of same words different meaning.

# Summary

| Algorithm | Performance |
|---|---|
| Cosine similarity | 0.4375 |
| BM25 | 0.5625 |
| Jaccard | 0.4417 |

Cosine distance takes the norm of the vectors into account and thus gets rid of the influence of the size of the vectors. So if the norm of vector is calculated it is better to use cosine.

Jaccard similarity performs reasonably well, normalized or not (by its definition, the length of the vectors are indirectly taken into account already...), although it looks that in my testing the cosine similarity is doing a little better. So this one is more convenient.

BM25 takes can solve the issue of document length by adjusting b, a little bit more steps but more efficient and accurate.

Thank you!