11-693 Software Methods for Biotechnology

Homework 3 report

Xuwei Zou
xuweiz

Task 1

1.1. Architecture of system

1.1.1. Type system

(1) Document type include four types: relevanceValue, queryID, text, tokenList. RelevanceValue indicates the rel number in each sentence, this value represent the type of this sentence. QueryId stores the number of query. Text records the sentences. TokenList stores the processed tokens.

(2) Token type include two types: text, frequency. Text stores each processed tokens. Frequency records how many times the tokens appear in this sentence.

1.1.2. Collection reader

Collection reader already provides in the archetype. In collection reader, all questions and documents will be read line by line and put into cases.

1.1.3. Analysis Engine

In analysis engine, sentences retrieved in collection reader will be separated into pieces of information and store in type system. These type system will be put back into cases.

1.1.4. CAS Consumer

In Cas Consumer, all questions will be store in a list and all document will be store in a map which the value will be a list stored the documents of one query. After all sentences are processed, all documents will be compared to its question and cosine similarity will be updated. Then, by using a comparator, the documents for the same question will be sorted. Also, MRR value will be calculated according the rank of the relevant documents. Finally, final report will be printed in the formation of given sample.

1.2. Final Performance

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/java (2014年
cosine=0.2791 rank=2 qid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash the Rom
cosine=0.2858 rank=2 qid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all time--made what
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 t
cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of Cl
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
cosine=0.1833 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably b
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 fe
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 t
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 ma
cosine=0.2265 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the systematized
cosine=0.1268 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burned and cra
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvan
(MRR) Mean Reciprocal Rank ::0.4375
Total time taken: 1.425
```

Task 2

2.1. Error analysis

In the 20 questions examples, there are three main reasons that cause failure of recognizing the correct document: tokens mismatch, too much irrelevant information, Different expression.

2.1.1. Tokens mismatch

This error may be caused by the difference in tenses/possessives/cases or punctuations attached between tokens in questions and correct documents.

Examples:

E1: qid=2 rel=99 What has been the largest crowd to ever come see Michael Jordan

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.

In this question, the token “Jordan--one” cannot be correctly recognized as Jordan.

E2: qid=5 rel=99 What river is called China's Sorrow?

cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrow of China"

In this question, several tokens failed to be correctly recognized due to cases different(e.g. “River,”⇔”river”) possessives different (e.g. “China” ⇔”China’s”) or tenses different(e.g. “call” ⇔”called”).

2.1.2. Too much irrelevant information

Correct document contained answer may also contains irrelevant information. These irrelevant information will impair the cosine similarity of the correct answer. Also, this may be caused by wrong documents not provide any answer to the question but use words sharing strong similarity

with the question.
Examples:
qid=1 rel=99 Give us the name of the volcano that destroyed the ancient city of Pompeii
cosine=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.
In this question, the correct document not only gives the correct answer (the name of the volcano) but also other irrelevant information (e.g. the time of eruption, casualty).

2.1.3. Different expression

Correct document provides correct answer that may not use the exact words in questions, or correct answer does not mentioned in the question.
Example:
qid=12 rel=99 What is the height of the tallest redwood?
cosine=0.3727 rank=4 qid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood stands 367 1/2 feet tall.
In this example, the query ask the height of the tallest redwood. Though the correct document gives the 367 1/2 feet answer, the document does not include the word “height” that exact match the question.

2.1.4. Main error type for each documents

		Questions
Tokens mismatch	punctuations attached	q2, q4, q6, q8, q20
	different tenses/possessives/cases	q3, q5, q7, c9, q11, q14, q16, q20
Too much irrelevant information		q1, q5, q7, q8, q17, q18, q19
Different expression		q12, q13, q15

2.2. Error handling

2.2.1. Tokens mismatch

For punctuations attached, I try to replace all punctuations with a blank.
In question 2, 4, 6, 8, 20:
Before replacement:

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/java (2014年10月21日 下午8:26:02)
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 t
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=0.5547 rank=2 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
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=0.3078 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) Mean Reciprocal Rank ::0.5000
Total time taken: 2.447
```

After replacement:

<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/java (2014年10月
cosine=0.3172 rank=1 qid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all time--made what was
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 N
cosine=0.6013 rank=1 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.2750 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inexplicably bit his
cosine=0.4104 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Pennsylvania is
(MRR) Mean Reciprocal Rank ::0.8000
Total time taken: 1.721

Conclusion:

Most results have great improvement, cosine similarity has raised.

For tenses/possessives/cases difference, I first replace all uppercase letter with lowercase letter, then use Stanford stemmer to unified tenses and possessives.

In question 3, 5, 7, 9, 11, 14, 16, 20:

Before modification:

cosine=0.2357 rank=3 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorrc
cosine=0.0891 rank=3 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that beca
cosine=0.5804 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.1768 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.4216 rank=2 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.2828 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Penr
(MRR) Mean Reciprocal Rank ::0.3854
Total time taken: 2.32

After modification:

cosine=0.4714 rank=1 qid=3 rel=1 Alaska was purchased from Russia in year 1867.
cosine=0.0990 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call "sorr
cosine=0.0891 rank=4 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that bec
cosine=0.5804 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.3536 rank=2 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.4216 rank=3 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.4243 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year, Per
(MRR) Mean Reciprocal Rank ::0.5521
Total time taken: 2.153

Conclusion:

Most results have improvement. However, since these documents may also affected by attached punctuations, some question may have a less effective of negative result. I will test to use both type modification.

For both kinds mismatch:

In question q2, q3, q4, q5, q6,q7, q8, q9,q11,q14,q16,q20

Before modification:

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin
cosine=0.2858 rank=2 qid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all tin
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
cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often c
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 tl
cosine=0.1833 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyso
cosine=0.5804 rank=2 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.1768 rank=4 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.4216 rank=2 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.2828 rank=3 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.3078 rank=2 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election y
(MRR) Mean Reciprocal Rank ::0.4236
Total time taken: 1.92
```

After modification:

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/jav
cosine=0.4282 rank=1 qid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all time--
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 gar
cosine=0.2970 rank=2 qid=5 rel=1 People of China have mixed feelings about River, which they often call "
cosine=0.6682 rank=1 qid=6 rel=1 Roger Bannister was the first to break the four-minute mile barrier.
cosine=0.2535 rank=2 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that
cosine=0.3667 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson in
cosine=0.6529 rank=1 qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
cosine=0.5303 rank=1 qid=11 rel=1 Devils Tower can be found in Crook County
cosine=0.5270 rank=1 qid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
cosine=0.4243 rank=1 qid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
cosine=0.4104 rank=1 qid=20 rel=1 They call it the Keystone State, and in this unpredictable election year,
(MRR) Mean Reciprocal Rank ::0.8750
Total time taken: 2.748
```

Conclusion:

After combining both kinds of modification, the result has a significant improvement. Most relevant documents are retrieved as first rank. Token mismatch related errors are resolved by these methods.

2.2.2. Too much irrelevant information

To solve the irrelevant information, I try to remove words with no exact meaning (stopwords) before compute the cosine similarity.

In question q1,q5, q7, q8, q17, q18, q19

Before modification:

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/ja
cosine=0.2791 rank=2 qid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic
cosine=0.0000 rank=3 qid=5 rel=1 People of China have mixed feelings about River, which they often call
cosine=0.0891 rank=3 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska tha
cosine=0.1833 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson i
cosine=0.1508 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal tempe
cosine=0.2265 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the
cosine=0.1268 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg bu
(MRR) Mean Reciprocal Rank ::0.4048
Total time taken: 1.69
```

After token modification:

```
<terminated> VectorSpaceRetrieval [Java Application] /home/krypton/luna/eclipse/jre/bin/java
cosine=0.2927 rank=2 qid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic ash
cosine=0.2970 rank=2 qid=5 rel=1 People of China have mixed feelings about River, which they often call "so
cosine=0.2535 rank=2 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska that be
cosine=0.3667 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson inex
cosine=0.3475 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal temperatu
cosine=0.2752 rank=2 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the syst
cosine=0.2758 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg burne
(MRR) Mean Reciprocal Rank ::0.4524
Total time taken: 1.562
```

After stopwords removing:

```
cosine=0.1667 rank=3 qid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volcanic
cosine=0.5000 rank=2 qid=5 rel=1 People of China have mixed feelings about River, which they often cal
cosine=0.5000 rank=1 qid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska tha
cosine=0.4781 rank=2 qid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 28, 1997, Tyson
cosine=0.2041 rank=3 qid=17 rel=1 Corn futures found support from forecasts for above-normal tempe
cosine=0.1132 rank=3 qid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948, the
cosine=0.1543 rank=3 qid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg b
(MRR) Mean Reciprocal Rank ::0.4762
Total time taken: 2.074
```

Conclusion:

For these questions, removing stopwords have a slightly bonus for performance. However, this modification also have negative influence for some cases. Moreover, for total 20 queries, implement removing stopwords actually lower the performance.

The low improvement may result by three reasons:

(1) Question for each query are rather short, and removing stopwords make it harder to match the correct answer.

(2) Removing stop words does not remove all irrelevant information in correct document.

(3) Removing stop words have no effects on the wrong documents sharing strong similarity with the queries.

2.2.3. Different expression

This kind of errors has the greatest difficulty to resolve. To find the link between words in question and document may solve this problem. In the example, “367 1/2 feet tall“ actually refer to “height” and this two words should be considered as same word. However, this linkage should consider about all kinds of condition in question (e.g. years, height, and place). For the time limitation, I have not implement any algorithm yet.