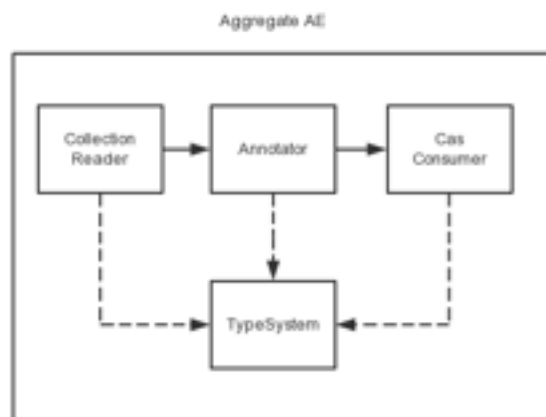


Hw3 Engineering and Error Analysis with UIMA

1 Task 1

The System Framework(showed in the figure below) is consist of four main components and they are implemented by the following java class.



Aggregate Analysis Engine: VectorSpaceRetrieval.java

CollectionReader: DocumentReader.java.

It reads file and splits each sentence into some parts in order to fill up the properties of document and update CAS.

Annotator: DocumentVectorAnnotator.java.

Here, I use HashMap to store the tokens and their frequency. By reading each document text, I compute the times that each token shows up. Then, I construct a vector of tokens and update the tokenList in CAS.

CasConsumer: RetrievalEvaluator.java.

This is the heart of the whole system, where I calculate cosine similarity between the query and its answer, find out the rank of the given answer sentence with relevance 1, and compute the MRR with all of the ranks.

Here are two main functions that is the processCas(), and collectionProcessComplete(). Because the pipeline processes one CAS at a time, I store data in memory as some private variable mem-

bers, like query id number list, text relevant values list, rankList storing ranks of each answer with relevance 1.

Here is a flow chart of logical design of processCas() method.



In collectionProcessComplete(), I calculate the final MRR of all the ranks in the rankList.

TypeSystem: Token.java, Document.java.

Results: MRR=0.43749999

cosine=0.2791	rank=2	aid=1	rel=1	In A.D. 79, long-dormant Mount <u>Yesuvius</u> erupted, burying in volc
cosine=0.2858	rank=2	aid=2	rel=1	When <u>Michael Jordan</u> --one of the greatest basketball player of al
cosine=0.2357	rank=3	aid=3	rel=1	Alaska was purchased from <u>Russia</u> in year 1867.
cosine=0.2315	rank=2	aid=4	rel=1	On March 2, 1962, <u>Wilt Chamberlain</u> scored a record 100 points ir
cosine=0.0000	rank=3	aid=5	rel=1	People of <u>China</u> have mixed feelings about <u>River</u> , which they off
cosine=0.5547	rank=2	aid=6	rel=1	<u>Roger Bannister</u> was the first to break the four-minute mile barr
cosine=0.0891	rank=3	aid=7	rel=1	And that's not even to mention the breathtaking beauty of Alaska
cosine=0.1833	rank=2	aid=8	rel=1	Fighting for Holyfield's WBA <u>heavyweight</u> title on June 28, 1997,
cosine=0.5804	rank=2	aid=9	rel=1	<u>Luna 2</u> was the first <u>spacecraft</u> to reach the surface of the Moor
cosine=0.5000	rank=1	aid=10	rel=1	<u>Menchu</u> won the <u>Nobel</u> peace prize in 1992.
cosine=0.1768	rank=4	aid=11	rel=1	<u>Devils Tower</u> can be found in Crook County
cosine=0.3162	rank=3	aid=12	rel=1	Named the " <u>Mendocino Tree</u> ," the 600- to 800-year-old <u>redwood</u> i
cosine=0.1195	rank=3	aid=13	rel=1	Oregon's Crater Lake tops it at 1,932 feet at its greatest depth
cosine=0.4216	rank=2	aid=14	rel=1	<u>Lionel Richie</u> was lead singer and <u>songwriter</u> for Commodores.
cosine=0.0788	rank=3	aid=15	rel=1	A new look at NASA satellite data revealed that Earth set a new
cosine=0.2828	rank=3	aid=16	rel=1	<u>Bob Marley</u> died in 1981 from cancer at age 36.
cosine=0.1508	rank=3	aid=17	rel=1	Corn futures found support from forecasts for above-normal tem
cosine=0.2265	rank=2	aid=18	rel=1	From a single hamburger stand in <u>San Bernardino, Calif.</u> , in 1944
cosine=0.1268	rank=3	aid=19	rel=1	On May 6, 1937, the hydrogen-filled German <u>dirigible</u> <u>Hindenburg</u>
cosine=0.3078	rank=2	aid=20	rel=1	They call it the <u>Keystone</u> State, and in this unpredictable elect
MRR=0.4374999999999999				

2 Task 2 Error Analysis

The following table is a summary of error types.

ID	Description	Error Type	Frequency
0	misspelling, like "Pompei", "Aaska"	misspelling	4
1	Too many tokens, describing in details	Long Length	9
2	punctuation missed or inserted, like A and A's are considered as different token	Punctuation mismatched	5
3	the different tense of one verb is considered differently	tense error	4
4	explain the same meaning with another description, like spaceship and spacecraft	vocabulary mismatch	10
5	Lowercase and uppercase mismatched	letter case error	4
6	function words like "a", "and", "the" help higher other less related answers' rank, thus lowering gold document's rank	stop words interference	5
7	singular and plural form	morphological variants	3

3. Task 2 Improvements Based on Error Analysis

Based on these types, I camp up with improvements from three aspects.

Here is a table for Improvements.

Note: All the improvements could be checked in my github repository. <https://github.com/eyrelzy/hw3-zhiyuel2>.

ID	Improvements	Improvement document('+' means getting higher rank, while '-' means lower its rank)	# of correct hit(baseline is 1)	MRR(baseline is 0.43)
	Solution			
1	Normalize a token , remove case sensitive(case conversion)	+3	2	0.4583
2	remove 's & s'	+2	3	0.4792
3	remove punctuation	+9	7	0.6125
4	introduce stem, Map a token to another token (solve morphological variants)	+4-2	8	0.6542
5	discard stop words	+1-5	9	0.6500
6	misspelling like more than one white spaces:Trim the text first, and also remove whose length is zero.	0	9	0.6500
7	Split up compound into several words	+1	10	0.675
8	First remove some punctuation, then do stemming method, and choose one of the scoring methods below			
8.0	cosine similarity	+2-1	11	0.7042
8.1	TF-IDF	+2-1	10	0.7125
8.2	BM25	+5-3	12	0.7823

1.Tokenize Improvement:

1) Normalize a token, case conversion:

By doing so, MRR increases by 0.4583, three items' rank increase, and even one successfully selects the most relevant answer.

For example:

qid=10 rel=99 Who won the **Nobel Peace Prize** in 1992?

qid=10 rel=1 Menchu won the **Nobel peace prize** in 1992.

After turning all the letters into lowercase, query 10 selected the gold answer.

```

1 cosine=0.2667 rank=2 aid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying in volce
2 cosine=0.3266 rank=1 aid=2 rel=1 When Michael Jordan--one of the greatest basketball player of all
3 cosine=0.3536 rank=2 aid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 cosine=0.2315 rank=2 aid=4 rel=1 On March 2, 1962, Wilt Chamberlain scored a record 100 points in
5 cosine=0.0990 rank=3 aid=5 rel=1 People of China have mixed feelings about River, which they ofte
6 cosine=0.5547 rank=2 aid=6 rel=1 Roger Bannister was the first to break the four-minute mile barri
7 cosine=0.0845 rank=4 aid=7 rel=1 And that's not even to mention the breathtaking beauty of Alaska
8 cosine=0.1833 rank=2 aid=8 rel=1 Fighting for Holyfield's NBA heavyweight title on June 28, 1997,
9 cosine=0.5804 rank=2 aid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of the Moon.
10 cosine=0.7500 rank=1 aid=10 rel=1 Menchu won the Nobel peace prize in 1992.
11 cosine=0.1768 rank=4 aid=11 rel=1 Devils Tower can be found in Crook County
12 cosine=0.3162 rank=4 aid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old redwood st
13 cosine=0.1195 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greatest depth.
14 cosine=0.4216 rank=3 aid=14 rel=1 Lionel Richiewas was lead singer and songwriter for Commodores.
15 cosine=0.0756 rank=3 aid=15 rel=1 A new look at NASA satellite data revealed that Earth set a new r
16 cosine=0.2828 rank=3 aid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.3015 rank=3 aid=17 rel=1 Corn futures found support from forecasts for above-normal temp
18 cosine=0.2831 rank=1 aid=18 rel=1 From a single hamburger stand in San Bernardino, Calif., in 1948,
19 cosine=0.2417 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Hindenburg
20 cosine=0.3078 rank=2 aid=20 rel=1 They call it the Keystone State, and in this unpredictable electi
21 MRR=0.47916666666666667

```

2) Remove 's and s' :

Increase two ranks, and hit three gold answers. However, this could also introduce some errors because of the part of speech difference for each word, for example, in general and general's have different meanings, and we should distinguish them.

i.e. qid=18 rel=99 Where was the first McDonald's built?

Removing "s" from helps generate a key word "mcdonald".

3) improve punctuation like "," in token "1823,":

It enhances 8 ranks, also hit 7 gold answers.

```
1 cosine=0.3112 rank=2 aid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius erupted, burying
2 cosine=0.3266 rank=1 aid=2 rel=1 When Michael Jordan--one of the greatest basketball play
3 cosine=0.3536 rank=2 aid=3 rel=1 Alaska was purchased from Russia in year 1867.
4 cosine=0.3886 rank=1 aid=4 rel=1 On March 2, 1962, Wilt Chamberlain scored a record 100 p
5 cosine=0.3862 rank=2 aid=5 rel=1 People of China have mixed feelings about River, which
6 cosine=0.5547 rank=2 aid=6 rel=1 Roger Bannister was the first to break the four-minute m
7 cosine=0.1690 rank=3 aid=7 rel=1 And that's not even to mention the breathtaking beauty o
8 cosine=0.2750 rank=2 aid=8 rel=1 Fighting for Holyfield's WBA heavyweight title on June 2
9 cosine=0.6529 rank=1 aid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of
10 cosine=0.8750 rank=1 aid=10 rel=1 Menchu won the Nobel peace prize in 1992.
11 cosine=0.1768 rank=4 aid=11 rel=1 Devils Tower can be found in Crook County
12 cosine=0.3953 rank=3 aid=12 rel=1 Named the "Mendocino Tree," the 600- to 800-year-old red
13 cosine=0.2390 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it at 1,932 feet at its greate
14 cosine=0.5270 rank=1 aid=14 rel=1 Lionel Richie was lead singer and songwriter for Comm
15 cosine=0.1512 rank=3 aid=15 rel=1 A new look at NASA satellite data revealed that Earth se
16 cosine=0.2828 rank=3 aid=16 rel=1 Bob Marley died in 1981 from cancer at age 36.
17 cosine=0.3769 rank=2 aid=17 rel=1 Corn futures found support from forecasts for above-norm
18 cosine=0.2831 rank=1 aid=18 rel=1 From a single hamburger stand in San Bernardino, Calif.,
19 cosine=0.2820 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-filled German dirigible Win
20 cosine=0.4104 rank=1 aid=20 rel=1 They call it the Keystone State, and in this unpredictab
MRR=0.6125
```

i.e. qid=18 rel=99 Where was the first McDonald's built?

Removing "?" from helps generate a key word "built".

4) misspelling like more than one white spaces:

Trim the text first, and also remove whose length is zero.

i.e. qid=19 rel=1 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.

There are more than one spaces between the words in red.

2. Stemming Improvement:

1) Given stemming method:

After using the given stemming method, it improves the MRR value due to correcting some morphological variants errors since it enables a broader range of queries to (correctly) match.

i.e. here purchase has morphological variants.

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.

After stemming, we have the following sentences.

in which year do a purchase of alaska happen?

alaska be purchase from russium in year 1867.

However, it also lower someone's rank, such as the 18 query. And the reason is this approach also helps other less related answer increase its cosine similarity.

What's more, I found out that the given method could not break down an entire compound as two tokens. i.e. "super_testers" will be considered as "super". Based on this, I made a slight change in the code by splitting up the compounds into separate tokens. (See the comment of deal with compound, to split up into two words in DocumentVectorAnnotator.java)

Another **disadvantage** is that some morphological variants should not be removed in order to avoid the misunderstanding of sentence. Here is an example.

- "Apple iPods are the new..." → Apple
- "Apple picking season has ..." → apple
- "Apples are popular food ..." → apple
- "Apple sales were up in May." → ? apple or Apple

Therefore, stemming is expected to improve when users are tolerant of stemming mistakes because there are so many relevant documents. What's more, a good stem must be context-dependent, but it is too expensive to do when processing many documents.

```

1 cosine=0.2667 rank=2 aid=1 rel=1 In A.D. 79, long-dormant Mount V
2 cosine=0.4003 rank=1 aid=2 rel=1 When Michael Jordan--one of the
3 cosine=0.4714 rank=1 aid=3 rel=1 Alaska was purchased from Russia
4 cosine=0.3086 rank=1 aid=4 rel=1 On March 2, 1962, Wilt Chamberla
5 cosine=0.3062 rank=2 aid=5 rel=1 People of China have mixed feeli
6 cosine=0.5547 rank=2 aid=6 rel=1 Roger Bannister was the first to
7 cosine=0.1782 rank=3 aid=7 rel=1 And that's not even to mention t
8 cosine=0.3667 rank=2 aid=8 rel=1 Fighting for Holyfield's WBA hea
9 cosine=0.6529 rank=1 aid=9 rel=1 Luna 2 was the first spacecraft
10 cosine=0.8750 rank=1 aid=10 rel=1 Menchu won the Nobel peace prize
11 cosine=0.3536 rank=2 aid=11 rel=1 Devils Tower can be found in Cro
12 cosine=0.3953 rank=4 aid=12 rel=1 Named the "Mendocino Tree," the
13 cosine=0.2390 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it at
14 cosine=0.5270 rank=1 aid=14 rel=1 Lionel Richiewas was lead singer
15 cosine=0.1455 rank=3 aid=15 rel=1 A new look at NASA satellite dat
16 cosine=0.4243 rank=1 aid=16 rel=1 Bob Marley died in 1981 from can
17 cosine=0.3769 rank=2 aid=17 rel=1 Corn futures found support from
18 cosine=0.2265 rank=2 aid=18 rel=1 From a single hamburger stand in
19 cosine=0.2820 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-fil
20 cosine=0.4104 rank=1 aid=20 rel=1 They call it the Keystone State,
21 MRR=0.6541666666666667
--

```

2) Remove stop words

Stop words: Words that are discarded from a document representation.

- Typically function words: a, an, and, as, for, in, of, the, to, ...

By removing stop words, we could reduce index size significantly thus improving retrieval accuracy.

Nevertheless, it also has disadvantages.

Discarding stop words makes some queries difficult to satisfy, like “What is the keystone state?”. If we remove the stop words, it leaves only keystone and state. In this case, if the gold answer is quite long, its cosine similarity must be small.

Some other examples like “let it be or not”, will even leave nothing to retrieve. You can see from

```

1 cosine=0.0737 rank=3 aid=1 rel=1 In A.D. 79, long-dormant Mount V
2 cosine=0.1048 rank=3 aid=2 rel=1 When Michael Jordan--one of the
3 cosine=0.5303 rank=1 aid=3 rel=1 Alaska was purchased from Russia
4 cosine=0.3563 rank=1 aid=4 rel=1 On March 2, 1962, Wilt Chamberla
5 cosine=0.3441 rank=1 aid=5 rel=1 People of China have mixed feeli
6 cosine=0.2097 rank=2 aid=6 rel=1 Roger Bannister was the first to
7 cosine=0.2357 rank=1 aid=7 rel=1 And that's not even to mention t
8 cosine=0.4339 rank=2 aid=8 rel=1 Fighting for Holyfield's NBA hea
9 cosine=0.2649 rank=1 aid=9 rel=1 Luna 2 was the first spacecraft
10 cosine=0.7906 rank=1 aid=10 rel=1 Menchu won the Nobel peace prize
11 cosine=0.2500 rank=4 aid=11 rel=1 Devils Tower can be found in Cro
12 cosine=0.1443 rank=4 aid=12 rel=1 Named the "Mendocino Tree," the
13 cosine=0.3086 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it at
14 cosine=0.5774 rank=1 aid=14 rel=1 Lionel Richie was lead singer
15 cosine=0.2222 rank=3 aid=15 rel=1 A new look at NASA satellite dat
16 cosine=0.5477 rank=1 aid=16 rel=1 Bob Marley died in 1981 from can
17 cosine=0.1907 rank=3 aid=17 rel=1 Corn futures found support from
18 cosine=0.0000 rank=2 aid=18 rel=1 From a single hamburger stand in
19 cosine=0.1066 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-fil
20 cosine=0.3244 rank=1 aid=20 rel=1 They call it the Keystone State,
21 MRR=0.6500000000000001

```

its results that the MRR value REDUCES to 0.6500.

An increasingly common solution to this problem is to discard stop words from queries and occasionally leave them in the query if they are more than half the query terms or if user indicates that they should be retained. I've implemented this rule in this system, results of the given dataset show no difference by doing so, but I believe it do works in a bigger and complicated query set.

Another thing is about the stop word lists, which is always created manually. And the retrieval results also depend on its quality.

As for the given stop word lists, we could see some words like which, what, when, and they frequently appears in the query. In this case, if we remove them according to the rules for stop words, we would misunderstand the meaning of the query document, in a broad sense, it will lead to bad retrieval results.

3) Decompound: Split up compound into several words

By doing so, it increase some cosine similarity value and even higher its rank. For example, there is term “Jordan—one” in the second query sentence, and this approach increases its cosine to about 0.02. And I also successfully divided the following compounds.

```

[long, dormant] | 2
[nba, record] | 2
[169, 147] | 2
[four, minute] | 2
[4, minute] | 2
[cool, think] | 2
[800, year, old] | 3

```

[370, foot, tall]13
 [singer, songwriter]12
 [above, normal]12
 [fast, food]12
 [hydrogen, fill]12
 [hindenburg, class]12

The MRR value reaches 0.675 after accumulate all the methods and tricks above.

```

1 cosine=0.0722 rank=3 aid=1 rel=1 In A.D. 79, long-dormant Mount
2 cosine=0.1069 rank=3 aid=2 rel=1 When Michael Jordan--one of t
3 cosine=0.4714 rank=1 aid=3 rel=1 Alaska was purchased from Rus
4 cosine=0.3563 rank=1 aid=4 rel=1 On March 2, 1962, Wilt Chambe
5 cosine=0.3441 rank=1 aid=5 rel=1 People of China have mixed fe
6 cosine=0.4041 rank=1 aid=6 rel=1 Roger Bonnister was the first
7 cosine=0.2357 rank=1 aid=7 rel=1 And that's not even to mentio
8 cosine=0.4339 rank=2 aid=8 rel=1 Fighting for Holyfield's WBA
9 cosine=0.6529 rank=1 aid=9 rel=1 Luna 2 was the first spacecra
10 cosine=0.7906 rank=1 aid=10 rel=1 Menchu won the Nobel peace pr
11 cosine=0.2500 rank=4 aid=11 rel=1 Devils Tower can be found in
12 cosine=0.3727 rank=4 aid=12 rel=1 Named the "Mendocino Tree," t
13 cosine=0.3086 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it
14 cosine=0.5270 rank=1 aid=14 rel=1 Lionel Richie was lead sin
15 cosine=0.1455 rank=3 aid=15 rel=1 A new look at NASA satellite
16 cosine=0.5477 rank=1 aid=16 rel=1 Bob Marley died in 1981 from
17 cosine=0.1865 rank=3 aid=17 rel=1 Corn futures found support fr
18 cosine=0.0000 rank=2 aid=18 rel=1 From a single hamburger stand
19 cosine=0.1054 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-
20 cosine=0.4104 rank=1 aid=20 rel=1 They call it the Keystone Sta
21 MRR=0.675
22

```

3. Better or different similarity measures:

1) Dice coefficient and Jaccard coefficient

Actually, they are quite similar to the cosine similarity. They are not metric.

2) unnormalized TF-IDF

As for the methods above, all terms are treated as equally important. TF-IDF introduces an idea of adding weight to some terms. Here is how we calculate normalized TF-IDF, and I implemented one removing the normalization. It is meaningful for some cases like the second query, since it increase its rank by increasing the weight of some query key words only show up in itself. Also, this unnormalized method partly solves over long length issue. And we could see a lot of gold answer document with a long length, which gives rise to its bad rank if using normalization.

- * "t": document term weight = $\log(\text{tf}) + 1$
- * "c": collection term weight = $\log(N / \text{df})$
- * "c": cosine length normalization
- * "n": weight = 1.0 (i.e., no normalization)
- * For example:

$$\text{Cosine similarity metric} = \frac{\sum d_i \cdot q_i}{\sqrt{\sum d_i^2} \cdot \sqrt{\sum q_i^2}} = \frac{\sum (\log(gf) + 1) \cdot \left((\log(qgf) + 1) \log \frac{N}{df} \right)}{\sqrt{\sum (\log(gf) + 1)^2} \cdot \sqrt{\sum \left((\log(qgf) + 1) \log \frac{N}{df} \right)^2}}$$

"doc length normalization" "user weights" "idf" "query length normalization"

Know this

However, we could see a lot of zero similarity values, which means it has no words included in the query document. In this case, this method could not be a suitable one for this dataset. What's more, results showed that the high MRR is due to using the trick that highest relevant document has rank higher than others with the same similarity.

Further looking into the intermediate results, I found that some answer could hit because all of the candidate document has the same tf-idf value and according to our trick about tie that if we have two document with the same similarity value, we give our gold answer with a higher rank. So, I believe this tf-idf method is somewhat overfit for this dataset.

```

1 cosine=0.0000 rank=2 aid=1 rel=1 In A.D. 79, long-c
2 cosine=0.5883 rank=2 aid=2 rel=1 When Michael Jorda
3 cosine=0.9389 rank=1 aid=3 rel=1 Alaska was purchas
4 cosine=0.0000 rank=1 aid=4 rel=1 On March 2, 1962,
5 cosine=0.0000 rank=1 aid=5 rel=1 People of China ha
6 cosine=0.0000 rank=1 aid=6 rel=1 Roger Bannister wa
7 cosine=0.0000 rank=1 aid=7 rel=1 And that's not eve
8 cosine=2.4039 rank=1 aid=8 rel=1 Fighting for Holyf
9 cosine=0.0000 rank=1 aid=9 rel=1 Luna 2 was the fir
10 cosine=0.0000 rank=1 aid=10 rel=1 Menchu won the Nob
11 cosine=0.2937 rank=3 aid=11 rel=1 Devils Tower can b
12 cosine=0.0000 rank=1 aid=12 rel=1 Named the "Mendoci
13 cosine=0.0000 rank=3 aid=13 rel=1 Oregon's Crater La
14 cosine=0.4592 rank=2 aid=14 rel=1 Lionel Richie was v
15 cosine=0.0000 rank=1 aid=15 rel=1 A new look at NAS
16 cosine=0.0000 rank=1 aid=16 rel=1 Bob Marley died ir
17 cosine=0.0000 rank=1 aid=17 rel=1 Corn futures found
18 cosine=0.0000 rank=1 aid=18 rel=1 From a single hamb
19 cosine=0.2865 rank=2 aid=19 rel=1 On May 6, 1937, th
20 cosine=1.7824 rank=1 aid=20 rel=1 They call it the #
21 MRR=0.8333333333333334

```

Based on this overfit issue, I slightly change the formula of calculating the tf-idf value by adding one to the idf value, and get a more reliable result by successfully removing some ties.

```

1 cosine=6.5785 rank=3 aid=1 rel=1 In A.D. 79, long-dormant Mount Vesuvius
2 cosine=12.8259 rank=1 aid=2 rel=1 When Michael Jordan--one of the greatest
3 cosine=8.0089 rank=2 aid=3 rel=1 Alaska was purchased from Russia in year
4 cosine=8.0905 rank=2 aid=4 rel=1 On March 2, 1962, Wilt Chamberlain score
5 cosine=3.0596 rank=1 aid=5 rel=1 People of China have mixed feelings abou
6 cosine=6.7418 rank=2 aid=6 rel=1 Roger Bannister was the first to break 4
7 cosine=3.0228 rank=1 aid=7 rel=1 And that's not even to mention the breat
8 cosine=6.0177 rank=1 aid=8 rel=1 Fighting for Holyfield's WBA heavyweight
9 cosine=5.7198 rank=2 aid=9 rel=1 Luna 2 was the first spacecraft to reach
10 cosine=5.3025 rank=1 aid=10 rel=1 Menchu won the Nobel peace prize in 199
11 cosine=-0.6625 rank=1 aid=11 rel=1 Devils Tower can be found in Crook Count
12 cosine=1.8722 rank=4 aid=12 rel=1 Named the "Mendocino Tree," the 600- to
13 cosine=0.2757 rank=1 aid=13 rel=1 Oregon's Crater Lake tops it at 1,932 fe
14 cosine=9.8539 rank=1 aid=14 rel=1 Lionel Richie was lead singer and soc
15 cosine=5.5407 rank=2 aid=15 rel=1 A new look at NASA satellite data reveal
16 cosine=0.4374 rank=2 aid=16 rel=1 Bob Marley died in 1981 from cancer at
17 cosine=3.1306 rank=3 aid=17 rel=1 Corn futures found support from forecast
18 cosine=0.2400 rank=1 aid=18 rel=1 From a single hamburger stand in San Ber
19 cosine=5.8633 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-filled Gerra
20 cosine=6.5224 rank=1 aid=20 rel=1 They call it the Keystone State, and in
21 MRR=0.7125

```

2)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). Here is its score metric.

After utilizing this approach, I obtained the following results.

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

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

Observing its results shows its IDF doesn't give a positive impact on the dataset.

1 cosine=-0.7263	rank=2	aid=1	rel=1	In A.D. 79, long-dormant Mount Vesuv
2 cosine=0.0000	rank=1	aid=2	rel=1	When Michael Jordan--one of the gre
3 cosine=-5.0515	rank=3	aid=3	rel=1	Alaska was purchased from Russia in
4 cosine=-13.3896	rank=3	aid=4	rel=1	On March 2, 1962, Wilt Chamberlain :
5 cosine=-7.6626	rank=3	aid=5	rel=1	People of China have mixed feelings
6 cosine=-9.6347	rank=3	aid=6	rel=1	Roger Bannister was the first to br
7 cosine=-2.8654	rank=4	aid=7	rel=1	And that's not even to mention the l
8 cosine=-7.6505	rank=4	aid=8	rel=1	Fighting for Holyfield's WBA heavyw
9 cosine=-8.5812	rank=4	aid=9	rel=1	Luna 2 was the first spacecraft to i
10 cosine=-6.7293	rank=4	aid=10	rel=1	Menchu won the Nobel peace prize in
11 cosine=-8.3741	rank=4	aid=11	rel=1	Devils Tower can be found in Crook
12 cosine=-9.0645	rank=4	aid=12	rel=1	Named the "Mendocino Tree," the 600-
13 cosine=-7.4488	rank=3	aid=13	rel=1	Oregon's Crater Lake tops it at 1,9:
14 cosine=-1.7855	rank=2	aid=14	rel=1	Lionel Richiewas was lead singer and
15 cosine=-2.6887	rank=1	aid=15	rel=1	A new look at NASA satellite data r
16 cosine=-8.5024	rank=3	aid=16	rel=1	Bob Marley died in 1981 from cancer
17 cosine=-4.5603	rank=3	aid=17	rel=1	Corn futures found support from for
18 cosine=-3.5445	rank=1	aid=18	rel=1	From a single hamburger stand in Sai
19 cosine=-4.5632	rank=3	aid=19	rel=1	On May 6, 1937, the hydrogen-filled
20 cosine=-7.3847	rank=3	aid=20	rel=1	They call it the Keystone State, and
21 MRR=0.48639455782312924				

Furthermore, since there are several methods to calculate the IDF, this inspired me I added one

cosine=0.9880	rank=1	qid=1	rel=1	In A.D. 79, long-dormant Mount Vesuvius eru
cosine=3.0961	rank=1	qid=2	rel=1	When Michael Jordan--one of the greatest ba
cosine=1.9885	rank=1	qid=3	rel=1	Alaska was purchased from Russia in year 18
cosine=-5.3182	rank=3	qid=4	rel=1	On March 2, 1962, Wilt Chamberlain scored a
cosine=-0.2501	rank=1	qid=5	rel=1	People of China have mixed feelings about
cosine=-2.4000	rank=3	qid=6	rel=1	Roger Bannister was the first to break the
cosine=2.2073	rank=1	qid=7	rel=1	And that's not even to mention the breathta
cosine=-0.6867	rank=1	qid=8	rel=1	Fighting for Holyfield's WBA heavyweight ti
cosine=-0.5629	rank=3	qid=9	rel=1	Luna 2 was the first spacecraft to reach th
cosine=2.7562	rank=1	qid=10	rel=1	Menchu won the Nobel peace prize in 1992.
cosine=-4.5629	rank=4	qid=11	rel=1	Devils Tower can be found in Crook County
cosine=-4.9391	rank=4	qid=12	rel=1	Named the "Mendocino Tree," the 600- to 800
cosine=-3.6209	rank=3	qid=13	rel=1	Oregon's Crater Lake tops it at 1,932 feet
cosine=7.8736	rank=1	qid=14	rel=1	Lionel Richiewas was lead singer and songwr
cosine=1.0584	rank=1	qid=15	rel=1	A new look at NASA satellite data revealed
cosine=-2.7088	rank=3	qid=16	rel=1	Bob Marley died in 1981 from cancer at age
cosine=-0.8478	rank=1	qid=17	rel=1	Corn futures found support from forecasts f
cosine=-1.7230	rank=1	qid=18	rel=1	From a single hamburger stand in San Bernar
cosine=-0.8483	rank=2	qid=19	rel=1	On May 6, 1937, the hydrogen-filled German
cosine=0.2053	rank=1	qid=20	rel=1	They call it the Keystone State, and in thi

MRR=0.7823129251700679

when calculating the IDF. After that, MRR reaches 0.7823, and we have hit 12 correct selected documents.

Why this have a good result?

1. every word has a weight
2. it considers the length of a document
3. tie rule

Since I still found out some pseudo good results based on our tie rules, I tried to slightly change the formula and makes it more reliable. What I found out is that the parameters in the formula is quite important to fit this dataset.

Here is the result of various parameters for the BM25 formula.

The Okapi BMxx model

$$\sum_{t \in q} \left(\log \frac{N - df_t + 0.5}{df_t + 0.5} \right) \frac{tf_t}{tf_t + k_1 \left((1-b) + b \frac{doclen}{avg_doclen} \right)} \frac{(k_3 + 1) qtf_t}{k_3 + qtf_t}$$

RSJ weight (idf)
tf weight
user weight

BMxx indicates different parameter settings

- Originally: $k_1=2$, $b=0.75$, $k_3=0$ (also used in Inquiry)
- BM25: $k_1=1.2$, $b=0.75$, $k_3=0-1000$
 $k_1=0.9$, $b=0.40$, $k_3=0-1000$ (large collections)*

	b=0.75
k1=0.9	0.7823
k1=1.2	0.7823
k1=2.0	0.7823

Although, I didn't see any change if slightly adjusting the parameters, I found out that the score changes, and the k1=2.0 makes the least number of ties.

3) cosine similarity

Finally, I returned to the cosine similarity approach. Here is the results after removing the unrelated punctuations and using stem function. There is almost no tie, and the results is more reliable.

```

1 cosine=0.1443 rank=3 aid=1 rel=1 In A.D. 79, long-dormant Maur
2 cosine=0.1543 rank=3 aid=2 rel=1 When Michael Jordan--one of 1
3 cosine=0.4714 rank=1 aid=3 rel=1 Alaska was purchased from Rus
4 cosine=0.4454 rank=1 aid=4 rel=1 On March 2, 1962, Wilt Chambe
5 cosine=0.5735 rank=1 aid=5 rel=1 People of China have mixed fu
6 cosine=0.4041 rank=1 aid=6 rel=1 Roger Bannister was the first
7 cosine=0.3354 rank=1 aid=7 rel=1 And that's not even to mentio
8 cosine=0.4339 rank=2 aid=8 rel=1 Fighting for Holyfield's WBA
9 cosine=0.6529 rank=1 aid=9 rel=1 Luna 2 was the first spacecra
10 cosine=0.7906 rank=1 aid=10 rel=1 Menchu won the Nobel peace pr
11 cosine=0.5000 rank=1 aid=11 rel=1 Devils Tower can be found in
12 cosine=0.3727 rank=4 aid=12 rel=1 Named the "Mendocino Tree," 1
13 cosine=0.3086 rank=3 aid=13 rel=1 Oregon's Crater Lake tops it
14 cosine=0.5270 rank=1 aid=14 rel=1 Lionel Richiewas was lead sir
15 cosine=0.1455 rank=3 aid=15 rel=1 A new look at NASA satellite
16 cosine=0.5477 rank=1 aid=16 rel=1 Bob Marley died in 1981 from
17 cosine=0.1865 rank=3 aid=17 rel=1 Corn futures found support fr
18 cosine=0.0778 rank=3 aid=18 rel=1 From a single hamburger stanc
19 cosine=0.1054 rank=3 aid=19 rel=1 On May 6, 1937, the hydrogen-
20 cosine=0.4104 rank=1 aid=20 rel=1 They call it the Keystone Stc
21 MRR=0.7041666666666666

```

4) evaluate the above three methods by p-value:

In statistical significance testing, the p-value is the probability of obtaining a test statistic result at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the p-value turns out to be less than a predetermined significance level, often 0.05. Such a result indicates that the observed result would be highly unlikely under the null hypothesis. It is quite similar to other statistical method like, student's t-test.

Here is a statement of how math lab compute the p-value. x,y is the vector of the score value o calculated by each method in the report file.

```
x=[6.5785, 12.8259, 8.0089,.....,6.5224]
```

```
[h,p,ci,stats] = ttest2(x,y)
```

From its result, I know that the p-value for every two methods is more than 0.5, so I could say that we could not know the reality of these results.

4.Follow-up Thoughts for Improvement:

1.Provide a inverted lists:

The above methods have not considered the relative position two words. For example, some disturbing candidate words have a good result only if they show up. However, they might in a meaningless order.

To improve this, we could store an relative position for each candidate token.

2. Morphological variants mismatched issue:

i.e.

```
qid=13      rel=99      How deep is Crater Lake?
```

```
qid=13      rel=1      Oregon's Crater Lake tops it at 1,932 feet at its  
greatest depth.
```

From this example, we could see that deep could not be consider the same as depth based on our methods. Because the stem method only care about some morphological variants like tense variants. And it could not recognize token like depth and deep as the same.

3. Vocabulary mismatched issue:

- 1) **paraphrase:** By scanning the datasets, I found out a lot of gold answers has tokens paraphrasing those in its query document.

For example.

```
qid=9 rel=99      What was the first spaceship on the moon
```

```
qid=9 rel=1 Luna 2 was the first spacecraft to reach the surface of  
the Moon.
```

However, our temporary approach could not recognize tokens with the same meaning. What's more the meaning needs a context-based solution, which has a great cost.

2) Phrase Recognition Methods: Part of Speech Tagging

Different words might have the same meaning, while one word could have multiple meaning, and also many part of speech.

By storing each word's part of speech information, we could partly put the token's context into consideration.

i.e. I/NN love/VV this/NN great/JJ city/NN.

4. Formula and parameters Impact

There are several interpretations for IDF and slight variations on its formula. Slightly changing its formula like use $\log(\text{term frequency})+1$, or just use term frequency leads to various results. Some results are overfit for this dataset because of the tie rules.

What's more, I noted that the formula for IDF that I used shows potentially major drawbacks since most of terms appearing in **more than half of the corpus documents**. These terms' IDF is negative, so for any two almost-identical documents, one which contains the term and one which does not contain it, the latter will possibly get a larger score. This means that terms appearing in more than half of the corpus will provide negative contributions to the final document score. This is often an undesirable behavior, so many real-world applications would deal with this IDF formula in a different way.

Again, in the calculation of BM25 derivation, the parameters of k_1 and b value don't make any change to the final result, but slightly changes each score.

5. Stem method introduce some drawbacks

With both stemming method and BM25, I have the following results if I didn't remove "" for the "devil" in each answer document. As you can see, it has four answers with the same score, and due to our tie rule, we make the the relevant one the highest.

i.e. devil tower can be find in crook county
to the west across the wyome border be the staggeringly beautiful devil'
il' tower national monument
devil' tower be an igneous intrusion that rise dramatically 1267 foot
(386 m) above the surround terrain
in 1941 petzoldt join other rock climber to rescue a maroon para-
chutist who have land atop devil' tower
what be the height of the tallest redwood
goldbm in gold query: -2.2809345280805604
answerbm in answer: -2.2809345280805604
answerbm in answer: -2.2809345280805604
answerbm in answer: -2.2809345280805604

After removing “”, MRR reduces because now every candidate could hit “devil”, and the each

1 cosine=0.9888	rank=1	qid=1	rel=1	In A.D. 79, long-dormant Mount Vesuvius
2 cosine=3.0961	rank=1	qid=2	rel=1	When Michael Jordan--one of the great
3 cosine=1.9885	rank=1	qid=3	rel=1	Alaska was purchased from Russia in 1867
4 cosine=-5.3182	rank=3	qid=4	rel=1	On March 2, 1962, Wilt Chamberlain scored
5 cosine=-0.2501	rank=1	qid=5	rel=1	People of China have mixed feelings about
6 cosine=-2.4000	rank=3	qid=6	rel=1	Roger Bannister was the first to break the
7 cosine=2.2073	rank=1	qid=7	rel=1	And that's not even to mention the brilliant
8 cosine=-0.6867	rank=1	qid=8	rel=1	Fighting for Holyfield's NBA heavyweight
9 cosine=-0.5629	rank=3	qid=9	rel=1	Luna 2 was the first spacecraft to reach
10 cosine=2.7562	rank=1	qid=10	rel=1	Menchu won the Nobel peace prize in 1938
11 cosine=-4.5629	rank=4	qid=11	rel=1	Devils Tower can be found in Crook County
12 cosine=-4.9391	rank=4	qid=12	rel=1	Named the "Mendocino Tree," the 600-foot
13 cosine=-3.6209	rank=3	qid=13	rel=1	Oregon's Crater Lake tops it at 1,937 feet
14 cosine=7.8736	rank=1	qid=14	rel=1	Lionel Richie was lead singer and
15 cosine=1.0584	rank=1	qid=15	rel=1	A new look at NASA satellite data reveals
16 cosine=-2.7088	rank=3	qid=16	rel=1	Bob Marley died in 1981 from cancer caused
17 cosine=-0.8478	rank=1	qid=17	rel=1	Corn futures found support from forecasts
18 cosine=-1.7230	rank=1	qid=18	rel=1	From a single hamburger stand in San Francisco
19 cosine=-0.8483	rank=2	qid=19	rel=1	On May 6, 1937, the hydrogen-filled Goodyear
20 cosine=0.2053	rank=1	qid=20	rel=1	They call it the Keystone State, and it's
1 MRR=0.7823129251700679				

score slightly changes. But, I believe this result is more reliable for a common dataset.

6. misspelling

There are some errors because of misspelling.

i.e. qid=7 rel=0 **Aaska** is a U.S. state situated in the northwest extremity of the North American continent.

Based on this, I tried to use lingpipe's spelling corrector. I used the given model, and found out it would not have a better performance for the misspelling words in our document. Here is the results. I also tried to add training documents into its dataset in order to obtain a new model. However, it still doesn't improve.

```
>Alaska
Found 32 document(s) that matched query 'Alaska':
Found 32 document(s) matching best alt='alaska':

>aaska
Found 0 document(s) that matched query 'aaska':
Found 0 document(s) matching best alt='was a':

>pompei
Found 0 document(s) that matched query 'pompei':
Found 1125 document(s) matching best alt='hockey':

>pompeii
Found 4 document(s) that matched query 'pompeii':
No spelling correction found.

>allaska
Found 0 document(s) that matched query 'allaska':
Found 32 document(s) matching best alt='alaska':
```

You could check my github link <https://github.com/eyrelzy/hw3-zhiyuel> to see how ling-pipe's corrector works.

7. normalization and unnormalized

I found out an interesting intermediate results that although our gold answer has more words matched, but still got a low rank because of over long length. In this case, we could consider other ways instead of normalization, like Pivoted Document Length Normalization.

Pivoted Document Length Normalization: Lnu.Ltu

Lnu.Ltu: A pivoted vector space similarity metric

- “L”: document term weight = $\frac{\log(tf) + 1}{\log(\text{avg } tf \text{ in doc}) + 1}$
- “t”: collection term weight = $\log(N / df)$
- “u”: pivoted unique normalization = $\frac{1}{0.8 \times + 0.2 \times \frac{\text{Num Unique Terms}}{\text{Avg Num Unique Terms}}}$
- “n”: weight = 1.0 (i.e., no normalization)
- For example:

$$\frac{\sum d_i \cdot q_i}{\sqrt{\sum d_i^2} \cdot \sqrt{\sum q_i^2}} = \frac{\sum (L_{d_i} \times n_{d_i}) \cdot (L_{q_i} \times t_{q_i})}{u_{d_i} \cdot u_{q_i}}$$

“tf”
“idf”

“doc length normalization”
“query length normalization”

8. These mentioned methods all have not thought of the meaning of the query and its answers.

To improve this, we need some methods based on the context. For example, if the query asked for “when”, we could give a bigger weight for answer containing “in” or numbers like “1967”. However, although it will have a good result for those queries that are asking something, this method is likely to be overfit to those dataset.

5 Reference

1. BM25:

http://en.wikipedia.org/wiki/Okapi_BM25

2. LingPipe spelling corrector

<http://alias-i.com/lingpipe/demos/tutorial/querySpellChecker/read-me.html>

3. p-value

<http://en.wikipedia.org/wiki/P-value>

4. TF-IDF

11-442 / 11-642 search engine courseware