CSC 483

Final Project

Yunxiao Hu

Dec. 8th, 2020

How to run the code:

Run the code in terminal.

(1) Get the source code in GitHub:

https://github.com/donshawhu/CSC483-Project.git

(2) Download the index folder at Google Drive:

https://drive.google.com/drive/folders/13CcMa9DaOAgV5gcHPcRNFP7O5XWcvoSa?usp=sharing , which name is “haveLemmaNoStemm” in the index folder. This is the best situation for my project.

(3) Put the index folder in the source code folder. The path should be created as: ./index/ haveLemmaNoStemm.

(4) Using terminal to locate the source code folder.

(5) Then, typing the following:

mvn compile

mvn test

Then, the score for the best situation (Lemmatization but no Stemming) will display in two ways:

Default similarity (BM25), Boolean similarity and Classic Similarity (Implementation of TFIDF similarity after Luence 6) and MutiSimilarity (Bool and Classic Similarity)

The result can be like: (I tested on mac and another pc locally, so it can have no problem to run)

```
MacBook-Pro:yunxiaohu_csc483_watson yunxiaohu$ mvn compile
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< 483:yunxiaohu_csc483_watson >-----------
---------
[INFO] Building yunxiaohu_csc483_watson 0.0.1-SNAPSHOT
[INFO] --------------------------------[ jar ]----------------------
---------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @
yunxiaohu_csc483_watson ---
[INFO] Using 'UTF-8' encoding to copy filtered resources.
[INFO] skip non existing resourceDirectory /Users/yunxiaohu/eclipse-
workspace/yunxiaohu_csc483_watson/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @
yunxiaohu_csc483_watson ---
[INFO] Nothing to compile - all classes are up to date
[INFO] ------------------------------------------------------------------
---------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------
---------
[INFO] Total time:  0.697 s
[INFO] Finished at: 2020-12-08T02:39:10-07:00
[INFO] ------------------------------------------------------------------
---------
MacBook-Pro:yunxiaohu_csc483_watson yunxiaohu$ mvn test
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< 483:yunxiaohu_csc483_watson >-----------
---------
[INFO] Building yunxiaohu_csc483_watson 0.0.1-SNAPSHOT
[INFO] --------------------------------[ jar ]----------------------
---------
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @
yunxiaohu_csc483_watson ---
[INFO] Using 'UTF-8' encoding to copy filtered resources.
```

```
[INFO] skip non existing resourceDirectory /Users/yunxiaohu/eclipse-
workspace/yunxiaohu_csc483_watson/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @
yunxiaohu_csc483_watson ---
[INFO] Nothing to compile - all classes are up to date
[INFO]
[INFO] --- maven-resources-plugin:2.6:testResources (default-
testResources) @ yunxiaohu_csc483_watson ---
[INFO] Using 'UTF-8' encoding to copy filtered resources.
[INFO] skip non existing resourceDirectory /Users/yunxiaohu/eclipse-
workspace/yunxiaohu_csc483_watson/src/test/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:testCompile (default-testCompile)
@ yunxiaohu_csc483_watson ---
[INFO] Nothing to compile - all classes are up to date
[INFO]
[INFO] --- maven-surefire-plugin:2.22.0:test (default-test) @
yunxiaohu_csc483_watson ---
[INFO]
[INFO] -------------------------------------------------------
[INFO]  T E S T S
[INFO] -------------------------------------------------------
[INFO] Running yunxiaohu_csc483_watson.LemmaTest
=============================================================
Running similarity Default(BM25).
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS
tagger from edu/stanford/nlp/models/pos-tagger/english-left3words-
distsim.tagger ... done [0.5 sec].
Hit List Position 9 find: Taiwan    answer: Taiwan (influence MMR
score)
Hit List Position 8 find: Tintoretto    answer: Tintoretto (influence
MMR score)
Hit List Position 10 find: The Help    answer: The Help (influence MMR
score)
Hit List Position 2 find: Boot Hill    answer: Boot Hill (influence
MMR score)
Hit List Position 3 find: Knights Templar    answer: Knights Templar
(influence MMR score)
Hit List Position 9 find: Aaron Burr    answer: Aaron Burr (influence
MMR score)
Hit List Position 5 find: Michelle Obama    answer: Michelle Obama
(influence MMR score)
Hit List Position 9 find: Helsinki    answer: Helsinki (influence MMR
score)
Hit List Position 5 find: Hasbro    answer: Hasbro (influence MMR
score)
Hit List Position 3 find: Anna Paquin    answer: Anna Paquin
(influence MMR score)
```

```
Hit List Position 5 find: Henry Kissinger    answer: Henry Kissinger
(influence MMR score)
Hit List Position 10 find: Henry Wadsworth Longfellow    answer: Henry
Wadsworth Longfellow (influence MMR score)
Hit List Position 6 find: The Faerie Queene    answer: The Faerie
Queene (influence MMR score)
Hit List Position 3 find: Kangaroo    answer: Kangaroo (influence MMR
score)
Hit List Position 4 find: William Henry Harrison    answer: William
Henry Harrison (influence MMR score)
Hit List Position 2 find: Monrovia    answer: Monrovia (influence MMR
score)
Hit List Position 10 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 5 find: Rosa Parks    answer: Rosa Parks (influence
MMR score)
Hit List Position 5 find: Procter & Gamble    answer: Procter & Gamble
(influence MMR score)
Hit List Position 8 find: Khmer language    answer: Khmer language
(influence MMR score)
Hit List Position 7 find: Three's Company    answer: Three's Company
(influence MMR score)
Using the Default(BM25) Model:
    P@1: 20/100 = 0.2
    MMR: 0.24442857142857138

============================================================
Running similarity CLASSIC.
Hit List Position 6 find: Mayim Bialik    answer: Mayim Bialik
(influence MMR score)
Hit List Position 8 find: Duce    answer: Duce (influence MMR score)
Using the CLASSIC Model:
    P@1: 0/100 = 0.0
    MMR: 0.0029166666666666664

============================================================
Running similarity Bool.
Hit List Position 2 find: The Wall Street Journal    answer: The Wall
Street Journal (influence MMR score)
Hit List Position 8 find: Komodo dragon    answer: Komodo dragon
(influence MMR score)
Hit List Position 2 find: Ben Affleck    answer: Ben Affleck
(influence MMR score)
Hit List Position 3 find: Knights Templar    answer: Knights Templar
(influence MMR score)
Hit List Position 3 find: Michelle Obama    answer: Michelle Obama
(influence MMR score)
Hit List Position 2 find: Anna Paquin    answer: Anna Paquin
(influence MMR score)
```

```
Hit List Position 2 find: Henry Kissinger    answer: Henry Kissinger
(influence MMR score)
Hit List Position 4 find: Ouzo    answer: Ouzo (influence MMR score)
Hit List Position 3 find: Kangaroo    answer: Kangaroo (influence MMR
score)
Hit List Position 3 find: William Henry Harrison    answer: William
Henry Harrison (influence MMR score)
Hit List Position 5 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 2 find: Rosa Parks    answer: Rosa Parks (influence
MMR score)
Hit List Position 10 find: Harrison Ford    answer: Harrison Ford
(influence MMR score)
Using the Bool Model:
    P@1: 13/100 = 0.13
    MMR: 0.17508333333333334

============================================================
Running similarity MultiSimilarity.
Hit List Position 6 find: Taiwan    answer: Taiwan (influence MMR
score)
Hit List Position 10 find: Jackie Joyner-Kersee    answer: Jackie
Joyner-Kersee (influence MMR score)
Hit List Position 2 find: Knights Templar    answer: Knights Templar
(influence MMR score)
Hit List Position 5 find: Casablanca    answer: Casablanca (influence
MMR score)
Hit List Position 10 find: Aaron Burr    answer: Aaron Burr (influence
MMR score)
Hit List Position 2 find: Michelle Obama    answer: Michelle Obama
(influence MMR score)
Hit List Position 7 find: Helsinki    answer: Helsinki (influence MMR
score)
Hit List Position 2 find: Anna Paquin    answer: Anna Paquin
(influence MMR score)
Hit List Position 2 find: Henry Wadsworth Longfellow    answer: Henry
Wadsworth Longfellow (influence MMR score)
Hit List Position 3 find: William Henry Harrison    answer: William
Henry Harrison (influence MMR score)
Hit List Position 7 find: The Atlanta Journal-Constitution    answer:
The Atlanta Journal-Constitution (influence MMR score)
Hit List Position 3 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 2 find: Rosa Parks    answer: Rosa Parks (influence
MMR score)
Hit List Position 2 find: Rob Reiner    answer: Rob Reiner (influence
MMR score)
Hit List Position 4 find: Martin Sheen    answer: Martin Sheen
(influence MMR score)
```

```
Hit List Position 3 find: Janet Jackson    answer: Janet Jackson
(influence MMR score)
Hit List Position 3 find: Procter & Gamble    answer: Procter & Gamble
(influence MMR score)
Hit List Position 7 find: Heather Locklear    answer: Heather Locklear
(influence MMR score)
Using the MultiSimilarity:
    P@1: 20/100 = 0.2
    MMR: 0.2557857142857142

[INFO] Tests run: 1, Failures: 0, Errors: 0, Skipped: 0, Time elapsed:
10.893 s - in yunxiaohu_csc483_watson.LemmaTest
[INFO]
[INFO] Results:
[INFO]
[INFO] Tests run: 1, Failures: 0, Errors: 0, Skipped: 0
[INFO]
[INFO] ------------------------------------------------------------------
---------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------
---------
[INFO] Total time:  12.636 s
[INFO] Finished at: 2020-12-08T02:39:25-07:00
[INFO] ------------------------------------------------------------------
---------
MacBook-Pro:yunxiaohu_csc483_watson yunxiaohu$
```

Code Part

Totally has 5 classes.

Index.java:

Have main function to call CreateIndex and QueryEngine to create the indexes and get the score. There also have 4 Boolean to set the situation. Such as, Lemmatization but no Stemming, no Lemmatization but Stemming, no Lemmatization and no Stemming, create indexes or not, do the search or not.

Setting files:

String queries: Questions file passes to QueryEngine

String indexPath: The created index location by CreateIndex, needs to be used in CreateInlex and QueryEngine

String input_dir: The folder of wiki files, can be used in CreateIndex

Setting ways of creating index: (This answers the question 1 how to index the file)

boolean isLemma: Passes to CreateIndex and QueryEngine, controls whether uses Lemmatization in question query and index files.

boolean isStem: Passes to CreateIndex and QueryEngine, controls whether uses Stemming in question query and index files.

(isLemma and isStem cannot both equal to true) (Question 1)

Setting Index.java functions:

boolean isCreateIndex: Controls whether runs CreateIndex to create index.

boolean isRunQueries: Controls whether runs QueryEngine to search the questions.

CreateIndex.java:

Answer question 1 many points.

public void readFiles(String directory)

public void readFile(String filename, IndexWriter w)

These two functions can read all wiki files and index them and can show how to deal with every wiki pages. (Question 1)

private String cleanTextContent(String text)

This function is used in readFile(String filename, IndexWriter w), can clean all non-English text, which can improve the speed when index all files. (Question 1)

QueryEngine.java:

public String runQueries(String filename)

public List<ResultClass> runQuery(String query)

These two functions can read question file and do the search and calculate the score. (Question 1, Question 2, Question 3)

public String removeStopWords(String text)

This function can remove the stop words in query, which can improve a little the accuracy of searching. (Question 2)

LemmaOrStemm.java:

This class can lemmatization or stemming the input string by using StanfordNLP Sentence class. (Question 1)

ResultClass.java:

A container. Same as previous homework.

Question 1 Indexing and Retrieval

Indexing:

I prepared the terms for indexing by 3 situations (Lemmatization but no Stemming, no Lemmatization but Stemming, no Lemmatization no Stemming).

I found some wiki pages have some non-English content, so I remove those non-English content. I didn't remove the stop words when indexing because I thought it may reduce the speed of indexing files. But actually, if I remove the stop words when indexing, it can improve the accuracy of searching. It can be discussed in Question 5. When indexing the wiki pages, I found some wiki pages have some equal signs, so I remove them. And Some content between to those equal signs is useless such as "See also, Further Reading", so I remove them. I just divided the WIKI pages by title and content. Title format is [[Title]], but sometimes have like "[[File: ….]]". So, I do a double check to make sure get the right title.

Retrieval:

I used all words (category and clue) to create the query. However, I also lemmatized or stemming the query base on different situations. What's more, I removed the stop words in query.

Question 2 Measuring Performance

For this project, I used precision at 1, and MMR. Using P@1 because we count the number of the correct answers by top 1 hit list to see what proportion in whole 100 questions. The reason why I use MMR that it can tell us how much potential the system has for improvement. Because MMR is useful when we care about the correct answers position of the top 10 answers. Thus, according to MMR score we can improve the indexing way and using which similarity that can make all answers at the top 1 position

of hit list. The best situation is when MMR score equals to P@1 score which means all answers are at the first position of hit list. Thus, the accuracy can be improved.

Question 3 Changing the score function:

I change the default scoring similarity (BM25) by Classic Similarity (Implementation of TFIDF after Lucene 6) and Bool similarity, the result shows on the table below:

Situations	Default (BM25)		Classic Similarity (TFIDF)		Bool Similarity		MultiSimilarity (Bool Similarity and ClassicSimilarity)	
Only Lemmatization	P@1	0.2	P@1	0	P@1	0.13	P@1	0.2
	MMR	0.24	MMR	0.0029	MMR	0.175	MMR	0.255
Only Stemming	P@1	0.1	P@1	0	P@1	0.09	P@1	0.12
	MMR	0.15	MMR	0.0044	MMR	0.12	MMR	0.19
Do Nothing	P@1	0.2	P@1	0	P@1	0.1	P@1	0.17
	MMR	0.2502	MMR	0.0087	MMR	0.16	MMR	0.258

From this table, the default MultiSimilarity method has best score in P@1 and MMR metrics of all situations. It has the best performance. The Default Similarity (BM25) is second best performance similarity. And Classic Similarity is worst performance similarity.

Before I got these results, I predicated that BM25 would be best one, the second one can be Classic Similarity (Traditional TFIDF), Bool Similarity can be the worst one. That is because BM25 and TFIDF both are based on Vector Space Model and Bool

Model. And BM25 is improved by TFIDF. According to this condition, BM25 can have best performance but it would not be better too much than TFIDF. Thus, I did not understand why Classic Similarity had worst performance. And its MMR scores are also lowest one which means it does not have much improving potential. According to this situation, I used the 4th Similarity (MultiSimilarity) which combined BoolSimilarity and ClassicSimilarity and got the similar P@1 and MMR score. Thus, I guess that the combination of BoolSimilarity and ClassicSimilarity is the real TFIDF similarity. Furthermore, I also guess ClassicSimilarity in Lucene may not use Bool model which causes its performance not good as I predicated.

Question 4 Error Analysis:

In my best configuration Only Lemmatization, correctly / incorrectly in:

Default (BM25) is 20/80

Classic Similarity (TFIDF) is 0/100

Bool Similarity is 13/87

MultiSimilarity is 20/80.

In my opinion, the reason why the simple system can find the correct answers is some long clues have many relevance terms to the WIKI pages and I used category in clues which the WIKI pages also have categories, thus, they can help system to find some correct answers. What's more, I delete some non-English contents when indexing, it can help compress the length of the pages which can help increase the TF score to make the system find the correct pages.

Problems: (1) Short clue, (2) did not remove stop words and [tlp] staff (meaningless staff) when indexing, (3) Common terms, (4) Similarity have potential to improve causes of MMR scores.

Short clue:

Such as “GOLDEN GLOBE WINNERS In 2009: Sookie Stackhouse”, the answer Anna Paquin appears at the top 10 Hit list position 3, in my Only Lemmatization configuration Default model (BM25).

```
Hit List Position 3 find: Anna Paquin    answer: Anna Paquin (influence MMR score)
```

Fortunately, this situation can be solved easily by improving the similarity because it was found in the top 10 Hit list. However, some short clue answers do not appear in the top 10 Hit list, such as “NAME THE PARENT COMPANY Fisher-Price toys” which answer is “Mattel”. This situation is hard for me to improve the performance of the system. Even if I change the indexing ways of WIKI pages because it has few terms to get the right relevance page.

Did not remove stop words and [tlp] staff (meaning less staff):

Common terms:

These two classes can be talked together. Some common terms in clues like numbers or dates or some common words which may also appears in [tlp] staff. Thus, in this situation, it can affect the weight of some non-relevance pages to make them into the top 10 Hit list. Did not remove stop words when indexing can make the length of the pages longer which can cause the term frequency smaller to let the correct Wiki page escape from the system.

Similarity have potential to improve causes of MMR scores

From the output of my system, we can see there are some answers at top 10 Hit list, but they do not at the first one of the top 10 Hit lists. If we can improve the similarity, these correct answers can have chance to appear at the first one of the top 10 Hit lists to improve the performance of the system.

Best Configuration Analysis:

My best configuration is Only Lemmatization because all of the similarities in this configuration find the most correct answer in all 3 configurations. Only Stemming can make the system have worst performance. The reason I guess is stemming cuts off the end or the beginning of the word can make many terms be the same shapes. In this situation it hard for mine choices of similarity. It can make many terms scores in the clue be the same, so, it can reduce the top 10 Hit list accuracy.

The Output Of 3 Configurations in My Eclipse Console:

No Lemmatization or Stemming:

```
Running similarity Default(BM25).
Hit List Position 7 find: Tintoretto    answer: Tintoretto
(influence MMR score)
Hit List Position 2 find: Cairo    answer: Cairo (influence MMR
score)
Hit List Position 7 find: The Help    answer: The Help
(influence MMR score)
Hit List Position 2 find: Knights Templar    answer: Knights
Templar (influence MMR score)
Hit List Position 6 find: Casablanca    answer: Casablanca
(influence MMR score)
Hit List Position 4 find: Michelle Obama    answer: Michelle
Obama (influence MMR score)
```

```
Hit List Position 6 find: Hasbro    answer: Hasbro (influence
MMR score)
Hit List Position 3 find: Anna Paquin    answer: Anna Paquin
(influence MMR score)
Hit List Position 6 find: Henry Kissinger    answer: Henry
Kissinger (influence MMR score)
Hit List Position 3 find: James Dean    answer: James Dean
(influence MMR score)
Hit List Position 3 find: Mattel    answer: Mattel (influence
MMR score)
Hit List Position 4 find: Henry Wadsworth Longfellow    answer:
Henry Wadsworth Longfellow (influence MMR score)
Hit List Position 10 find: Cape Town    answer: Cape Town
(influence MMR score)
Hit List Position 7 find: The Faerie Queene    answer: The
Faerie Queene (influence MMR score)
Hit List Position 3 find: George Martin    answer: George Martin
(influence MMR score)
Hit List Position 6 find: William Henry Harrison    answer:
William Henry Harrison (influence MMR score)
Hit List Position 2 find: Monrovia    answer: Monrovia
(influence MMR score)
Hit List Position 2 find: Three's Company    answer: Three's
Company (influence MMR score)
Using the Default Model:
     P@1: 20/100 = 0.2
     MMR: 0.2502857142857143

Running similarity CLASSIC.
Hit List Position 6 find: Mayim Bialik    answer: Mayim Bialik
(influence MMR score)
Hit List Position 5 find: Feta    answer: Feta (influence MMR
score)
Hit List Position 2 find: Duce    answer: Duce (influence MMR
score)
Using the CLASSIC Model:
     P@1: 0/100 = 0.0
     MMR: 0.008666666666666666

Running similarity Bool.
Hit List Position 5 find: Tintoretto    answer: Tintoretto
(influence MMR score)
Hit List Position 9 find: Rotary International    answer: Rotary
International (influence MMR score)
Hit List Position 5 find: Heath Ledger    answer: Heath Ledger
(influence MMR score)
```

Hit List Position 3 find: Knights Templar answer: Knights Templar (influence MMR score)
Hit List Position 4 find: Aaron Burr answer: Aaron Burr (influence MMR score)
Hit List Position 2 find: Michelle Obama answer: Michelle Obama (influence MMR score)
Hit List Position 5 find: Hasbro answer: Hasbro (influence MMR score)
Hit List Position 2 find: Anna Paquin answer: Anna Paquin (influence MMR score)
Hit List Position 3 find: Henry Kissinger answer: Henry Kissinger (influence MMR score)
Hit List Position 5 find: James Dean answer: James Dean (influence MMR score)
Hit List Position 7 find: Henry Wadsworth Longfellow answer: Henry Wadsworth Longfellow (influence MMR score)
Hit List Position 6 find: Cape Town answer: Cape Town (influence MMR score)
Hit List Position 2 find: Governor General of Canada answer: Governor General of Canada (influence MMR score)
Hit List Position 5 find: Kangaroo answer: Kangaroo (influence MMR score)
Hit List Position 2 find: Feta answer: Feta (influence MMR score)
Hit List Position 3 find: William Henry Harrison answer: William Henry Harrison (influence MMR score)
Hit List Position 3 find: Alec Baldwin answer: Alec Baldwin (influence MMR score)
Hit List Position 2 find: Rosa Parks answer: Rosa Parks (influence MMR score)
Hit List Position 8 find: Procter & Gamble answer: Procter & Gamble (influence MMR score)
Hit List Position 4 find: Souvlaki answer: Souvlaki (influence MMR score)
Hit List Position 2 find: Robert Downey, Jr. answer: Robert Downey, Jr. (influence MMR score)
Using the Bool Model:
     P@1: 10/100 = 0.1
     MMR: 0.16378968253968254

Running similarity MultiSimilarity.
Hit List Position 5 find: Taiwan answer: Taiwan (influence MMR score)
Hit List Position 9 find: Rotary International answer: Rotary International (influence MMR score)

Hit List Position 3 find: Cairo answer: Cairo (influence MMR score)
Hit List Position 4 find: Heath Ledger answer: Heath Ledger (influence MMR score)
Hit List Position 4 find: George Michael answer: George Michael (influence MMR score)
Hit List Position 3 find: Knights Templar answer: Knights Templar (influence MMR score)
Hit List Position 8 find: Casablanca answer: Casablanca (influence MMR score)
Hit List Position 3 find: Aaron Burr answer: Aaron Burr (influence MMR score)
Hit List Position 2 find: Hasbro answer: Hasbro (influence MMR score)
Hit List Position 2 find: Anna Paquin answer: Anna Paquin (influence MMR score)
Hit List Position 7 find: Edinburgh answer: Edinburgh (influence MMR score)
Hit List Position 2 find: James Dean answer: James Dean (influence MMR score)
Hit List Position 4 find: Mattel answer: Mattel (influence MMR score)
Hit List Position 2 find: Henry Wadsworth Longfellow answer: Henry Wadsworth Longfellow (influence MMR score)
Hit List Position 3 find: Game Change answer: Game Change (influence MMR score)
Hit List Position 3 find: Cape Town answer: Cape Town (influence MMR score)
Hit List Position 2 find: Ouzo answer: Ouzo (influence MMR score)
Hit List Position 2 find: Kangaroo answer: Kangaroo (influence MMR score)
Hit List Position 2 find: William Henry Harrison answer: William Henry Harrison (influence MMR score)
Hit List Position 2 find: The Atlanta Journal-Constitution answer: The Atlanta Journal-Constitution (influence MMR score)
Hit List Position 2 find: Rosa Parks answer: Rosa Parks (influence MMR score)
Hit List Position 10 find: Harrison Ford answer: Harrison Ford (influence MMR score)
Hit List Position 9 find: Budapest answer: Budapest (influence MMR score)
Hit List Position 3 find: Rob Reiner answer: Rob Reiner (influence MMR score)
Hit List Position 7 find: Janet Jackson answer: Janet Jackson (influence MMR score)

```
Hit List Position 3 find: Procter & Gamble    answer: Procter &
Gamble (influence MMR score)
Hit List Position 6 find: Rickshaw    answer: Rickshaw
(influence MMR score)
Hit List Position 7 find: Robert Downey, Jr.    answer: Robert
Downey, Jr. (influence MMR score)
Using the MultiSimilarity:
     P@1: 17/100 = 0.17
     MMR: 0.2582579365079365
```

Only Lemmatization:

```
Running similarity Default(BM25).
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/Users/yunxiaohu/cs/483/hw/hw4/stanford-corenlp-
4.2.0/slf4j-simple.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in
[jar:file:/Users/yunxiaohu/.m2/repository/org/slf4j/slf4j-
simple/1.7.25/slf4j-simple-
1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for
an explanation.
SLF4J: Actual binding is of type
[org.slf4j.impl.SimpleLoggerFactory]
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger -
Loading POS tagger from edu/stanford/nlp/models/pos-
tagger/english-left3words-distsim.tagger ... done [0.5 sec].
Hit List Position 9 find: Taiwan    answer: Taiwan (influence
MMR score)
Hit List Position 8 find: Tintoretto    answer: Tintoretto
(influence MMR score)
Hit List Position 10 find: The Help    answer: The Help
(influence MMR score)
Hit List Position 2 find: Boot Hill    answer: Boot Hill
(influence MMR score)
Hit List Position 3 find: Knights Templar    answer: Knights
Templar (influence MMR score)
Hit List Position 9 find: Aaron Burr    answer: Aaron Burr
(influence MMR score)
Hit List Position 5 find: Michelle Obama    answer: Michelle
Obama (influence MMR score)
Hit List Position 9 find: Helsinki    answer: Helsinki
(influence MMR score)
```

```
Hit List Position 5 find: Hasbro    answer: Hasbro (influence
MMR score)
Hit List Position 3 find: Anna Paquin    answer: Anna Paquin
(influence MMR score)
Hit List Position 5 find: Henry Kissinger    answer: Henry
Kissinger (influence MMR score)
Hit List Position 10 find: Henry Wadsworth Longfellow    answer:
Henry Wadsworth Longfellow (influence MMR score)
Hit List Position 6 find: The Faerie Queene    answer: The
Faerie Queene (influence MMR score)
Hit List Position 3 find: Kangaroo    answer: Kangaroo
(influence MMR score)
Hit List Position 4 find: William Henry Harrison    answer:
William Henry Harrison (influence MMR score)
Hit List Position 2 find: Monrovia    answer: Monrovia
(influence MMR score)
Hit List Position 10 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 5 find: Rosa Parks    answer: Rosa Parks
(influence MMR score)
Hit List Position 5 find: Procter & Gamble    answer: Procter &
Gamble (influence MMR score)
Hit List Position 8 find: Khmer language    answer: Khmer
language (influence MMR score)
Hit List Position 7 find: Three's Company    answer: Three's
Company (influence MMR score)
Using the Default Model:
     P@1: 20/100 = 0.2
     MMR: 0.24442857142857138

Running similarity CLASSIC.
Hit List Position 6 find: Mayim Bialik    answer: Mayim Bialik
(influence MMR score)
Hit List Position 8 find: Duce    answer: Duce (influence MMR
score)
Using the CLASSIC Model:
     P@1: 0/100 = 0.0
     MMR: 0.0029166666666666664

Running similarity Bool.
Hit List Position 2 find: The Wall Street Journal    answer: The
Wall Street Journal (influence MMR score)
Hit List Position 8 find: Komodo dragon    answer: Komodo dragon
(influence MMR score)
Hit List Position 2 find: Ben Affleck    answer: Ben Affleck
(influence MMR score)
```

Hit List Position 3 find: Knights Templar answer: Knights Templar (influence MMR score)
Hit List Position 3 find: Michelle Obama answer: Michelle Obama (influence MMR score)
Hit List Position 2 find: Anna Paquin answer: Anna Paquin (influence MMR score)
Hit List Position 2 find: Henry Kissinger answer: Henry Kissinger (influence MMR score)
Hit List Position 4 find: Ouzo answer: Ouzo (influence MMR score)
Hit List Position 3 find: Kangaroo answer: Kangaroo (influence MMR score)
Hit List Position 3 find: William Henry Harrison answer: William Henry Harrison (influence MMR score)
Hit List Position 5 find: Alec Baldwin answer: Alec Baldwin (influence MMR score)
Hit List Position 2 find: Rosa Parks answer: Rosa Parks (influence MMR score)
Hit List Position 10 find: Harrison Ford answer: Harrison Ford (influence MMR score)
Using the Bool Model:
     P@1: 13/100 = 0.13
     MMR: 0.17508333333333334

Running similarity MultiSimilarity.
Hit List Position 6 find: Taiwan answer: Taiwan (influence MMR score)
Hit List Position 10 find: Jackie Joyner-Kersee answer: Jackie Joyner-Kersee (influence MMR score)
Hit List Position 2 find: Knights Templar answer: Knights Templar (influence MMR score)
Hit List Position 5 find: Casablanca answer: Casablanca (influence MMR score)
Hit List Position 10 find: Aaron Burr answer: Aaron Burr (influence MMR score)
Hit List Position 2 find: Michelle Obama answer: Michelle Obama (influence MMR score)
Hit List Position 7 find: Helsinki answer: Helsinki (influence MMR score)
Hit List Position 2 find: Anna Paquin answer: Anna Paquin (influence MMR score)
Hit List Position 2 find: Henry Wadsworth Longfellow answer: Henry Wadsworth Longfellow (influence MMR score)
Hit List Position 3 find: William Henry Harrison answer: William Henry Harrison (influence MMR score)

```
Hit List Position 7 find: The Atlanta Journal-Constitution
answer: The Atlanta Journal-Constitution (influence MMR score)
Hit List Position 3 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 2 find: Rosa Parks    answer: Rosa Parks
(influence MMR score)
Hit List Position 2 find: Rob Reiner    answer: Rob Reiner
(influence MMR score)
Hit List Position 4 find: Martin Sheen    answer: Martin Sheen
(influence MMR score)
Hit List Position 3 find: Janet Jackson    answer: Janet Jackson
(influence MMR score)
Hit List Position 3 find: Procter & Gamble    answer: Procter &
Gamble (influence MMR score)
Hit List Position 7 find: Heather Locklear    answer: Heather
Locklear (influence MMR score)
Using the MultiSimilarity:
     P@1: 20/100 = 0.2
     MMR: 0.2557857142857142
```

Only Stemming:

```
Running similarity Default(BM25).
Hit List Position 7 find: Tintoretto    answer: Tintoretto
(influence MMR score)
Hit List Position 2 find: Cairo    answer: Cairo (influence MMR
score)
Hit List Position 4 find: The Help    answer: The Help
(influence MMR score)
Hit List Position 2 find: Knights Templar    answer: Knights
Templar (influence MMR score)
Hit List Position 3 find: Michelle Obama    answer: Michelle
Obama (influence MMR score)
Hit List Position 6 find: Henry Kissinger    answer: Henry
Kissinger (influence MMR score)
Hit List Position 3 find: James Dean    answer: James Dean
(influence MMR score)
Hit List Position 2 find: The Faerie Queene    answer: The
Faerie Queene (influence MMR score)
Hit List Position 3 find: Lord Byron    answer: Lord Byron
(influence MMR score)
Hit List Position 6 find: Governor General of Canada    answer:
Governor General of Canada (influence MMR score)
```

```
Hit List Position 2 find: William Henry Harrison    answer:
William Henry Harrison (influence MMR score)
Hit List Position 6 find: Monrovia    answer: Monrovia
(influence MMR score)
Hit List Position 5 find: Rob Reiner    answer: Rob Reiner
(influence MMR score)
Hit List Position 2 find: Three's Company    answer: Three's
Company (influence MMR score)
Hit List Position 2 find: Heather Locklear    answer: Heather
Locklear (influence MMR score)
Using the Default Model:
     P@1: 10/100 = 0.1
     MMR: 0.1509285714285714

Running similarity CLASSIC.
Hit List Position 9 find: Mayim Bialik    answer: Mayim Bialik
(influence MMR score)
Hit List Position 3 find: Duce    answer: Duce (influence MMR
score)
Using the CLASSIC Model:
     P@1: 0/100 = 0.0
     MMR: 0.0044444444444444444

Running similarity Bool.
Hit List Position 4 find: Tintoretto    answer: Tintoretto
(influence MMR score)
Hit List Position 2 find: Knights of Columbus    answer: Knights
of Columbus (influence MMR score)
Hit List Position 8 find: Ben Affleck    answer: Ben Affleck
(influence MMR score)
Hit List Position 10 find: Knights Templar    answer: Knights
Templar (influence MMR score)
Hit List Position 4 find: Michelle Obama    answer: Michelle
Obama (influence MMR score)
Hit List Position 2 find: Hasbro    answer: Hasbro (influence
MMR score)
Hit List Position 3 find: James Dean    answer: James Dean
(influence MMR score)
Hit List Position 9 find: Game Change    answer: Game Change
(influence MMR score)
Hit List Position 2 find: William Henry Harrison    answer:
William Henry Harrison (influence MMR score)
Hit List Position 5 find: Alec Baldwin    answer: Alec Baldwin
(influence MMR score)
Hit List Position 9 find: Rob Reiner    answer: Rob Reiner
(influence MMR score)
```

Hit List Position 8 find: Ottoman Empire   answer: Ottoman Empire (influence MMR score)
Hit List Position 10 find: Robert Downey, Jr.   answer: Robert Downey, Jr. (influence MMR score)
Using the Bool Model:
     P@1: 9/100 = 0.09
     MMR: 0.12205555555555554

Running similarity MultiSimilarity.
Hit List Position 3 find: The Wall Street Journal   answer: The Wall Street Journal (influence MMR score)
Hit List Position 2 find: Tintoretto   answer: Tintoretto (influence MMR score)
Hit List Position 2 find: Cairo   answer: Cairo (influence MMR score)
Hit List Position 3 find: Knights of Columbus   answer: Knights of Columbus (influence MMR score)
Hit List Position 2 find: Ben Affleck   answer: Ben Affleck (influence MMR score)
Hit List Position 2 find: Knights Templar   answer: Knights Templar (influence MMR score)
Hit List Position 2 find: Michelle Obama   answer: Michelle Obama (influence MMR score)
Hit List Position 2 find: Hasbro   answer: Hasbro (influence MMR score)
Hit List Position 3 find: The Faerie Queene   answer: The Faerie Queene (influence MMR score)
Hit List Position 3 find: Lord Byron   answer: Lord Byron (influence MMR score)
Hit List Position 2 find: Ouzo   answer: Ouzo (influence MMR score)
Hit List Position 8 find: Governor General of Canada   answer: Governor General of Canada (influence MMR score)
Hit List Position 8 find: George Martin   answer: George Martin (influence MMR score)
Hit List Position 2 find: William Henry Harrison   answer: William Henry Harrison (influence MMR score)
Hit List Position 2 find: The Atlanta Journal-Constitution   answer: The Atlanta Journal-Constitution (influence MMR score)
Hit List Position 3 find: Alec Baldwin   answer: Alec Baldwin (influence MMR score)
Hit List Position 3 find: Janet Jackson   answer: Janet Jackson (influence MMR score)
Hit List Position 10 find: Three's Company   answer: Three's Company (influence MMR score)

Hit List Position 7 find: Heather Locklear answer: Heather Locklear (influence MMR score)
Using the MultiSimilarity:
    P@1: 12/100 = 0.12
    MMR: 0.18992857142857145