Q1

Describe how you prepared the terms for indexing (stemming,lemmatization, stop words, etc.).

I prepared the wiki file by lemmatization. I used multi thread to add space between special symbols and “[]” in order to separate the word when doing the search. And turn the word with all capital letters into lower case. and ture plural into singular.

What issues specific to Wikipedia content did you discover, and how did you address them?

I found that Wikipedia content is complicate so I used multi thread to maximum utilization of CPU

Describe how you built the query from the clue.

I used two sublists for clues, a tittleList stores all categories and contentList stores contents for corresponding categories.

Are you using the category of the question?

yes.

Q2

```
#########################################
######  Lemma: true Stem: true           ######
######  MRR: 0.2893981467266319      ######
######  Correct: 22 / 100                  ######
######  Within 2~100: 46 / 100           ######
#########################################
```

Q3

tf-idf score used

```
#########################################
######  Lemma: true Stem: true           ######
######  MRR: 0.029045679757041924   ######
######  Correct: 2 / 100                    ######
######  Within 2~100: 19 / 100          ######
########################################
```

Using the tf-idf score decreased the correct number and the Within 2~100 number.

Q4

How many questions were answered correctly/incorrectly?

Correct: 22 / 100

Why do you think the correct questions can be answered by such a simple system?

because when we have a massive amount of data, it will be easier to get the right answer.

What problems do you observe for the questions answered incorrectly?

I found that questions answered incorrectly have one common feature: they all have longer categories so they will have more content that needs to be matched, and that will cause incorrect answers.

eg1. OLD YEAR'S RESOLUTIONS
content:old content:year content:resolut

eg2. NOTES FROM THE CAMPAIGN TRAIL
content:note content:from content:campaign content:trail

Lastly, what is the impact of stemming and lemmatization on your system?

```
########################################
######  Lemma: true Stem: true          ######
######  MRR: 0.2893981467266319        ######
######  Correct: 22 / 100                ######
######  Within 2~100: 46 / 100          ######
########################################

########################################
######  Lemma: false Stem: false        ######
######  MRR: 0.2333419866729001        ######
######  Correct: 16 / 100               ######
######  Within 2~100: 50 / 100         ######
########################################

########################################
######  Lemma: true Stem: false          ######
######  MRR: 0.25431424081629134      ######
######  Correct: 19 / 100                ######
######  Within 2~100: 44 / 100          ######
#######################################
```

```
##########################################
######  Lemma: false Stem: true            ######
######  MRR: 0.28658215789269825     ######
######  Correct: 22 / 100                     ######
######  Within 2~100: 48 / 100             ######
########################################
```

comparing the result of *Lemma: false* ***Stem: false*** and *Lemma: false* ***Stem: true*** we can see the correct number has a relatively larger change. So stem do help my system getting a better result, so is lemma but not more obvious than stem.

Q5

Adding tokenized for improving.

input string:
“British Standards are the standards produced by BSI Group which is incorporated under a Royal Charter (and which is formally designated as the National Standards Body (NSB) for the UK). The BSI Group produces British Standards under the authority of the Charter, which lays down as one of the BSI's objectives to: Formally, as per the 2002 Memorandum of Understanding between the BSI and the United Kingdom Government, British Standards are defined as:"

output:

```
[british] [standard] [standard] [produc] [bsi] [group] [which] [incorpor] [under] [royal] [charter]
[which] [formal] [design] [nation] [standard] [bodi] [nsb] [uk] [bsi] [group] [produc] [british]
[standard] [under] [author] [charter] [which] [lai] [down] [on] [bsi] [object] [formal] [per] [2002]
[memorandum] [understand] [between] [bsi] [unit] [kingdom] [govern] [british] [standard]
[defin]
```