11-791: Design and Engineering of Intelligent Information System

Homework 3

Name: Yulong Pei
Andrew ID: yulongp

1 Error Analysis

The sentences selected by the default cosine similarity are shown in Table 1. In fact, among all the 20 queries, only query 10 obtains the correct answer. For the other wrong selection, the reason for some representative queries are shown below.

Query id	questions	selected answers
1	Give us the name of the volcano that destroyed the ancient city of Pompeii	Vesuvius is located near the ruins of the destroyed city of Pompeii.
2	What has been the largest crowd to ever come see Michael Jordan	A supposedly last play of Michael Jordan gathered some of the largest crowd in history of NBA.
3	In which year did a purchase of Alaska happen?	William Seward negotiated a purchase c Alaska for $7.2 million.
4	What year did Wilt Chamberlain score 10(points?	A 100 point game was a highlight in a career of Wilt Chamberlain
5	What river is called China's Sorrow?	Yellow river is often called the mother c China
6	Who was the first person to run the mile ir less than four minutes	It is hard for humans to run the mile fast than in four minutes
7	What year did Alaska become a state?	Also as it did in 1959, the state of Alask was struggling with how to pay its bills.
8	When did Mike Tyson bite Holyfield's ear'	Tyson was choked up as he spoke of biting Evander Holyfield's ear.
9	What was the first spaceship on the moon	Eagle was the first manned spacecraft

		that reached the surface of the moon
10	**Who won the Nobel Peace Prize in 1992?**	**Menchu won the Nobel peace prize in 1992.**
11	Where is Devil's Tower	Devil's Tower is an igneous intrusion th rises dramatically 1,267 feet (386 m) above the surrounding terrain.
12	What is the height of the tallest redwood?	Mendocino Tree is the tallest redwood i the world.
13	How deep is Crater Lake?	Crater Lake is a caldera lake in the western United States.
14	Who was the lead singer for the Commodores	The Commodores originally came together from groups the Mystics and th Jays.
15	What is the coldest place on earth?	Oymyakon is the coldest place in Russia
16	When did Bob Marley die	Bob Marley was a Jamaican reggae singer-songwriter, musician, and guitari who did achieve international fame.
17	Which U.S. state is the leading corn producer?	The United States is the world's leading producer of corn
18	Where was the first McDonald's built?	McDonald's Corporation is the world's largest chain of hamburger fast food restaurants.
19	The Hindenburg disaster took place in 193 in which New Jersey town?	The Hindenburg disaster took place as the German passenger airship LZ 129 Hindenburg caught fire and was destroyed during its attempt to dock wit its mooring mast
20	What is the Keystone State?	Keystone Resort is the largest ski resort in Summit County located in Keystone Colorado.

Table 1: The list of selected answers for each question.

Query 3: the tokenization problem. The same stem of purchase and purchased cannot be captured.

Query 9: the synonym problem. The similar meaning of spaceship and spacecraft cannot be captured.

Query 11, 14: the stop word *is* and *the* makes the selected sentence have more overlapping words with the question. Remove stop words may solve this problem.

Query 12 and 13: the semantics of the questions are not captured can only the overlapping words are used. For example, when ask about the height, the answer may contain tall and when ask about deep, the answer may contain depth.

Overall, th errors mainly due to several reasons including tokenization/lemmatization problem, synonym problem, letter case problem, semantic problem. A brief statistics is shown in Table 2.

Error type	number
tokenization/lemmatization problem	6
synonym problem	5
letter case problem	1
semantic problem	7

Table 2: Brief statistics of error type and number.

2 Architecture

2.1 Type system

In this implementation, there are two types, i.e., Document and Token. Document stores the vector of a document (the question and the answer in this task) and Token stores the term text and the corresponding term frequency.

2.2 Annotator

The Annotator is implemented with the java file DocumentVectorAnnotator. In this Annotator, each of the Document is tokenized and put into the sparse term vector.

2.3 CAS consumer

In this implementation, CAS consumer plays the most important role. It contains storing sentence information, similarity calculation, answer retrieval, and evaluation.

3 Improvement

In order to improve the performance of the QA system, several methods have been compared in this study and the results are shown in Table 3.

Method	**MRR**
tf + cosine	0.4375
tf + stop word removal + cosine	0.4917

tf*idf + NLTK tokenization + cosine	0.6083
tf*idf + NLTK tokenization + stop word remov + cosine	0.6083

Table 3: The comparison between different combination.

The corresponding code for the last two methods, i.e., with NLTK tokenization and tf*idf can be found at https://github.com/yulongp/hw3-yulongp/tree/master/hw3-yulongp/src/main/resources/similarity/