IEMS308 Homework 4

Yujia Zhai

Executive summary:

In this study, a Q&A system is built by applying text mining techniques, which is able to give answer to four different pre-determined types of questions. The four types of questions are listed as below, where X, Y, and Z are replaceable keywords.

(1) Which companies went bankrupt in month X of year Y?
(2) GDP-related questions:
    a. What affects GDP?
    b. What percentage of drop or increase is associated with Z?
(3) Who is the CEO of company X?

The Q&A system is developed based on scraped articles from 2013 to 2014. The procedure of generating answers is in the following order: (1) Input question; (2) Analyze question; (3) Retrieve information from documents; (4) Analyze answer; (5) Give out answer.

In this study, the Q&A system built efficiently answers the input questions, and also successfully gives correct answers in all sample outputs. Namely, this Q&A system helps retrieve desired information with high efficiency and precision.

Methodology and Technique

The detail of each step is shown as below:

(1) Input question through the interface;
(2) Use cosine distance to quantify similarity between questions to determine the type of the question and the type of the answer desired.
(3) Extract keywords from the question by using *Regex* and *pos_tag* techniques.
(4) Index articles by *Elasticsearch* library to find articles desired efficiently; After finding the result, using *Regex* techniques in HW3 to discover entities or percentage point. For all candidate CEO or company names, using *Counter()* to count the appearance in the related article. Result with the highest count will be return.

By using *Elasticsearch*, two type of indices was implemented. When extracting CEO names and Company bankrupted record, I used sentences indices. Since those two types of questions are commonly informative, which means using one sentence is enough. For GDP type of question, I used articles indices. Since GDP reason are usually related to the core meaning of an article.

During implementation of the algorithm, I found extracting *GDP affect reasons* question is hard to achieve. Therefore, I read the articles extracted from elastic search and manually conclude them to several keywords: [Storm, Sequester, EUC program, appreciation, Lower fuel prices].

How to use Q&A system?

(1) Prerequisite
   a. Latest version of Elastic search running at port 9200.
   b. python3
   c. nltk libraries
   d. python elastic search client library

(2) How to Run
   a. Navigate to the folder where HW4.py locate.
   b. In cmd, run: python3 HW4.py
   c. Then you'll be able to use the Q&A system.

Business insights:

With this Q&A system built, answers to questions can be directly obtained. Several business insights can be inferred from this.

(1) The CEO and Bankruptcy types of questions significantly reduce the time cost to obtain answers to these types of questions. For example, without the Q&A system, to obtain the answer to a CEO question, one has to use search engine to find information internet page of a company, or go to the Wikipedia page of that company, and find the CEO information in those pages. For Bankruptcy questions, one has to find a bankruptcy database to obtain information of all bankrupted companies in the history. Then filter the time factor to get companies desired.
However, assume the Q&A system is deployed online, one can type the question in the search engine and click “search”, and instead of clicking those links in the results page one by one to find the answer, the Q&A system will directly give out the answer on the top of the results page, which is highly efficient and clear.

(2) The GDP type of questions makes further progress on this. More than just giving out some answer based on pure fact, the system even gives out analytic answers that require some analysis of data. This means that one can even get those analytic insight results simply by searching the question, without doing the analysis on their own. With enough information stored in the database, this Q&A system is able to replace some simple functions of analytic tools and databases, which is important in some business fields.