Shengjie Liu
IEMS 308
Professor Klabjan
Homework Assignment 4

Q&A System

Summary:

By using text mining technique, a Q&A System is developed based on the corpus from all of texts in 2013 and 2014. With this system, we can give answers quickly to four types of questions:1. which companies went bankrupt in month x of year y? 2. what affects gdp? 3. what percentage of drop or increase in gdp is associated with factor z? 4. What is the CEO of company X? With this QA system, we can have some business insight to support our decisions.

Pipeline for our Q&A system:

Technique used in system:

(1) using cosine distance to measure similarity among questions allows us to build our question classifier to decide which type of question is and determine the answer type.
(2) using function extract_entity in our code to find the Keywords in our question.
(3) using elasticsearch library to index our articles, which is Very useful to find the articles what we want.
(4) building three separate functions to answer each type of the Question. In each function, we define rules to determine Which article and sentence have the highest score to our question.

How to use our Q&A system?

(1) making sure that elasticsearch is properly installed in your python library and your own PC terminal. If you do not install this library in terminal, you request to use this library in python will be refused.
(2) making sure that all of necessary libraries in python are Installed such as NLTK, NUMPY, PANDAS,and so on.

(3) opening hw4_shengjieliu.ipynb in terminal and typing your Questions as the argument, then getting the answer.

Sample Output:

```
In [549]: answer("What is the CEO of Twitter?")
Out[549]: 'Jack Dorsey'

In [551]: answer("What is the CEO of Microsoft?")
Out[551]: 'Steve Ballmer'

In [552]: answer("What is the CEO of Facebook?")
Out[552]: 'Mark Zuckerberg'

In [554]: answer("Which company went bankrupt in September 2008?")
Out[554]: 'Lehman Brothers'

In [561]: answer("Which company went bankrupt in July 2013?")
Out[561]: 'Detroit'

In [572]: answer("Which company went bankrupt in October 2014?")
Out[572]: 'Reuters'

In [573]: answer("What factors have the most effect on GDP?")
Out[573]: 'Consumption, consumer spending, government spending, investment, imports, exports, foreign t
          rade'

In [574]: answer("what percentage is assoicated with foreign trade?")
Out[574]: '0.7 percent'

In [578]: answer("what percentage change in GDP results from government spending?")
Out[578]: '.6 percent'
```

Business insight:

(1) For the Q&A System part, if we can increase the size of our corpus and add function in our model, we can quickly get our answer for our problem, which can increase our business efficiency.
(2) For the output part, it is easy for us to know economic Condition for whole society or financial situation for some Companies.

The most useful part of the Q&A system is that we can get the information what we want quickly and precisely.