IEMS308 HW4 Q&A System

Taige Hong

March 2019

**Executive Summary**

BusinessInsider is a portal for business news, and their articles contain facts and views about the business world. We are using its articles from 2013 and 2014 as the knowledge base to answer business-related questions.

We are using the natural language processing packages to extract keywords in the questions and get into the article base to find out the corresponding answers related to these keywords.

The final deliverable is a function that is capable of answering questions of type: “Which company went bankrupt in March 2013?”, “What affects GDP?”, “What percentage of drop or increase is associated with this property?”, “Who is the CEO of company X?”

**Process**

The process is straightforward for this project:

Reading data -> Classifying question -> Find keywords related to the question -> Get most frequently appeared answer

The overall process for question 1, 3, 4 are similar, with some minor differences because of the difference itself in the nature of the question, while in question 2 we are utilizing the tfidf algorithm to find out the most relative terms.

We are using Elasticsearch because of the large number of article-term pairs, this tool will boost the search speed a lot.

Question 1: Which company went bankrupt in X of year Y?

To be considered as candidate, an article should at least contain one word from list(“bankrupt”, “bankruptcy”, “liqudate”, “declare”), the name of the month and also the year. After filtering out the first round, we then tokenize the article by sentence and find out the sentences with “bankrupt” in it. We use Regex expression to find company names as it will be too time-consuming to call the function in HW3. Then we are finding the term that appears the most time and use that as the response. The major concern in this process is that the Regex expression may not be able to accurately distinguish a name being a company name or not, thus introducing a lot of unrelated term in the response. To fight against this, I included codes to take out expanded stopwords and month names. It turns out the result is rather satisfactory.

Question 2: What affects GDP?

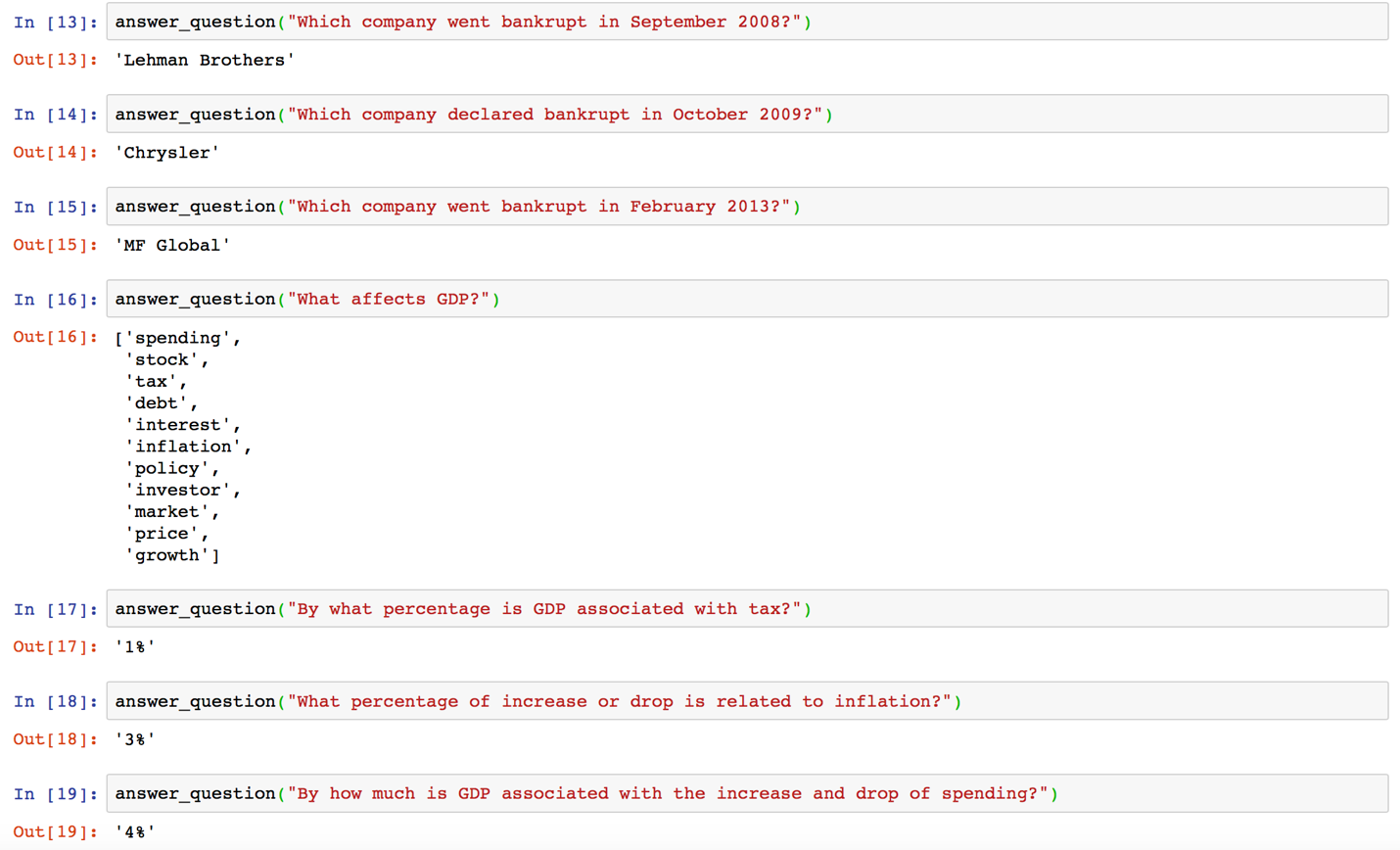
To answer this question, I used tfidf algorithm to select the words with highest score. The subset of articles are the ones containing “GDP” and “affect” or “effect”. The code spits out 50 best candidates and I manually chose the most meaningful ones as some of them are just spam words. There is indeed some way to improve this process, for example, expanding the stop word set or so, but it requires more time and work. As this question will not be asked in different forms, it is most efficient to do it this way as it will gives the most reasonable and accurate answers.

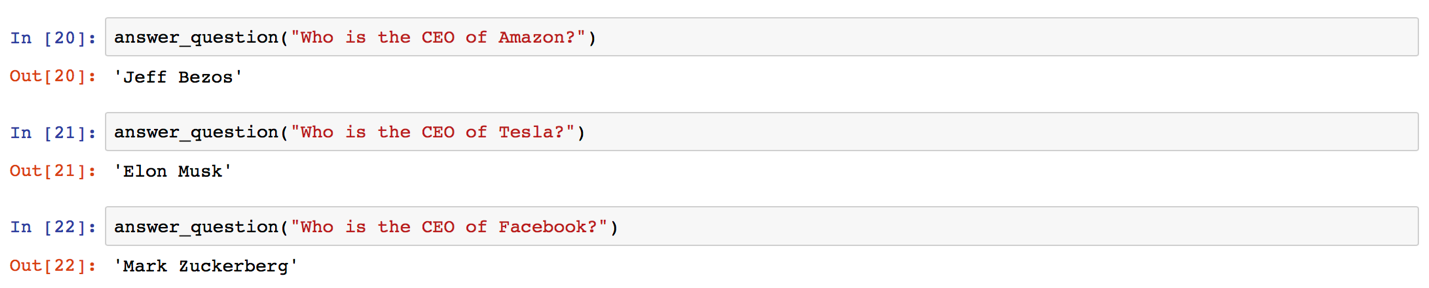
Question 3: What percentage of drop or increase is associated with this property?

The overall process is similar to question 1. The question itself has to include at least one word from the answer in question 2. Then it will filter the articles by “GDP” and the keyword chosen. Next step is tokenizing article into sentences and finding out ones with both “GDP” and the keyword in them. We then count the frequency of percentage (same Regex expression as in HW3) and use the one with highest frequency as the answer.

Question4: Who is the CEO of company X?  
We filter out articles with term “CEO” and the company name and use the same process on the corpus. The problem is sometimes the phrase will contain the company name and “CEO” in the front. To deal with it, I included codes to remove these terms from the phrase so that only name of CEO will be spit out.

**Result**

****



Overall the system is successful as the answers themselves are reasonable answers regarding to the questions asked. However, the answers may not be accurate. For example, Chrysler did actually file for bankruptcy, but in first half of 2009 instead of October 2009, and MF Global filed for bankruptcy in October 2011, not in February 2013. It turns out that the system is capable of capturing some of the companies going bankrupt, but the time does not always match. One of the reasons might be there are not many big companies going bankrupt every month, so the most related result might be some other bankrupt cases, spanning over months.

The CEO answer is quite successful, and the answer to the GDP factors is hard to assess as the question is not clear enough in the first place.

**Conclusion**

In this homework, we are able to develop a system that is able to extract keywords from questions and find the matching answers. It is a great way for us to get a view of the natural language processing system and to know about how it works behind the scene. There are many ways to improve, both in my program and in the overall NLP domain, thus it deserves more consideration and development.