**Q&A System Project4**

Hao Zhu

**Executive summary:**

In this case, we divide three required question in four different questions as “Which companies went bankrupt in month X of year Y?”, “What affects GDP?”, “What percentage of drop or increase is associated with this property?” and “Who is the CEO of company X?”. Different kinds of questions have different preprocess steps.

Then we use method of cosine similarity as question classifier, Elasticsearch as document selection and information retrieval. For the NER, we use the NLTK package because our own NER classifier do not perform well in project3.

When using this system, just input the question and then the program will give you the answer.

**Question Classifier:**

We set up four types of question and the question we raised as vector. Remove punctuation and do stemming. After that, we set up binary matrix, calculate the cosine similarity between the question we raised and other four origin types of question, get the question type that has the maximum cosine number.

**Question Processing:**

* **Bankrupt Question**: first we extract the month and year from the question and then set up query for search in Elasticsearch document:

Must have any of (bankrupt bankruptcy Bankrupt Bankruptcy) and Month and Year

Should have (article date has the month and year) and any of ['collapse', 'declared', 'protection', 'liquidat', 'failure', 'Chapter 11']

The “should” requirement is to improve the probability of correct answers. Since some of companies are rarely mentioned in the 730 articles.

Then get the top 20 scores articles. Choose the sentences that have the word “bankrupt”, then extract the company from those sentences. We choose the company name that counts the most as the final answer.

* **Factor of GDP**: Set up query for search in Elasticsearch document:

Must have [‘GDP gdp] and [‘affect’]

Then remove the punctuation, stop words and alpha word. Choose the sentence that at least have one ‘gdp’ and ‘affect’. Calculate the whole tf-idf of words matrix, get the words that have the largest tf-idf. After filtering, we get the factors that make sense.

* **Percentage Question**: first we extract the factors from the question and then set up query for search in Elasticsearch document:

Must have [‘GDP gdp] and [‘percent %’] and [factors] and ['increase', 'decrease', 'drop', 'growth', 'effect', 'affect']

Then get the top 20 scores articles. Choose the sentences that have the word “% or percent” and”gdp” and factor, then extract the percentage from those sentences. We choose the percent that counts the most as the final answer.

* **CEO Question** first we extract the company name from the question and then set up query for search in Elasticsearch document:

Must have [‘CEO ceo’] and [company name]

Then get the all articles that meet the requirements. Choose the sentences that have the word “ceo” and company name, then extract the PERSON from those sentences. We choose the name that counts the most as the final answer. In this question, we find that final list have large amount of candidates that include the name of companies, so we remove all those wrong entity and then we will get a perfect answer.

**Business Insight:**

This Q&A system can comprehensive analysis the factor of GDP, also the tf-idf method gives us another new view of how important each factor is recognized from the perspective of Finance News. We can then focus on the factor that has the largest tf-idf to see whether it is the one that most important on affecting GDP.

This Q&A system gives us a view of how severe and what the scale of a company’s bankruptcy. A company’s bankruptcy which is recognized by others as a big event will always be mentioned in the News after years. From that corpus, we know that Lehman Brothers went bankrupt in September 2008 is really a big event in the financial domain.

**Conclusion**

For these four questions, our system has different accuracy of answer.

The best performance of answer is for the question 4: answer CEO names. After remove the company names in the candidate list disturb the final answer, we can say the question 4’s answer is 100% correct in most of case.

The question 2 and question 3 is difficult to verify. Generally speaking, the factors we extract seems like correct. For the percentage, because we do not limit the year or other things, we can extract lots of related information. At last, we choose the one that is mentioned most fluently.

The worst performance of answer is for the question1. We change the document search query, number of documents chosen and type of NER classifier again and again, the accuracy of answer is changing in a wide range. The bigger companies are the larger the influence on our answer. Because the bankruptcy of larger companies will always be mentioned in different articles such as “Lehman Brothers”, also I find some National Institute related to bankruptcy also affect the results.