Executive Summary

A question and answer system is built based on Business Insider's news articles ranging from year 2013 to 2014. We uses techniques for information retrieval as well as natural language processing so that the system can automatically answer questions proposed by human beings within the closed-domain. More specifically, can answer four factual questions as followings:

1) Who is the CEO of a specific company?
2) Which company went bankrupt in a specific month of the X year?
3) What affects GDP?
4) What percentage of increase/decrease is related to the factor (above)

Taking a brief look at the result, first two questions can get very accurate answer while last two questions can return very vague/inaccurate result. The general process starts with establishing a document-term matrix into Elasticsearch database, classifying the input string as a specific type of question, forming a query based on the judgement, and then retrieve the document and get the answer. For the first two questions, since we filter through keywords to get candidate sentences and name entity recognition (NER) is used to filter through the answer, the result is very satisfying. On the other hand, there is no real "entity" that is responsible for GDP reasons, so keywords can be only extracted by tf-idf scores, which requires manually creates a stopword dictionary to get rid of some irrelevant words.

For business insights, it can clearly shown that rule-based approach for classification requires less work than training an automatic classifier. In addition, factual questions are easier to answer than conceptual questions since they only have one correct answer. Within the closed-domain, factual questions can be answered relatively quickly while the dataset might be biased.

Methodology

The detailed explanation for each step is described in the 308hw4deck.pdf. Business insights and further steps are also included there.

Instructions

1. Follow the instruction in the following website to Install Elasticsearch https://towardsdatascience.com/getting-started-with-elasticsearch-in-python-c3598e718380 (requires both zip file installation as well as Python module)
2. Get the jupyter notebook installed into the computer and change to the proper working directory
3. Use web browser to go to localhost:9200 to make sure the elasticsearch database is proper running
4. Run all cells so that all functions are implemented and database is install
5. Use the function "answer_the_question" and write the question as the parameter, the return value will be the answer of the question then.

Sample Output

```
print(answer_the_question("what impacts GDP"))
print(answer_the_question("what affects GDP"))
print(answer_the_question("what causes GDP"))
```

```
['public debt', 'hurricane', 'credit', 'market', 'cost', 'investment', 'tax', 'deficit', 'oil', 'spend', 'debt']
['public debt', 'hurricane', 'credit', 'market', 'cost', 'investment', 'tax', 'deficit', 'oil', 'spend', 'debt']
['public debt', 'hurricane', 'credit', 'market', 'cost', 'investment', 'tax', 'deficit', 'oil', 'spend', 'debt']
```

```
print(answer_the_question("what percent of increase in gdp is associated with oil price"))
print(answer_the_question("what percentage of drop in gdp is associated with tax"))
print(answer_the_question("what % of drop in gdp is caused by investment"))
```

```
0.2%
0.8%
1.18%
```

```
print (answer_the_question('who is the CEO of company Google?'))
print(answer_the_question('who is the CEO of Facebook'))
print(answer_the_question('who is the ceo of the company IBM'))
print(answer_the_question('who is the ceo of the company Morgan Stanley'))
```

```
Larry Page
Mark Zuckerberg
Ginni Rometty
James Gorman
```

```
print(answer_the_question('which comapany went bankrupt in July of 2013?'))
print(answer_the_question("Which company went bankrupt in September 2008?"))
print(answer_the_question("which company declared bankruptcy in June 2009?"))
```

```
Detroit
Lehman
GM
```