Question and Answer System

Executive Summary

The aim of the project was to create a Question and Answer System with the capability to answer the following types of questions:

- Which companies went bankrupt in month X of year Y?
- What affects GDP? What percentage of drop or increase is associated with this property?
- Who is the CEO of company X?

The solution was based on corpus of Business Insider articles from 2013-2014. ElasticSearch was utilized throughout the system for indexing, fast retrieval of documents and document scoring. Other steps were implemented using components taken from nltk and sklearn packages.
The system does well in extracting and providing the top sentence from the corpus based on the factual question types like:

```
Question:  Who is the CEO of Google?
{'CEO': 1, 'Google': 1}
I think the answer might be in: Larry Page (CEO of Google) Net Worth: $32.3
billion 4.
```

```
Question:  Which companies went bankrupt in month September of year 2008?
{'company': 1, 'companies': 1, 'go': 1, 'went': 1, 'bankrupt': 1, 'month': 1,
'September': 1, 'year': 1, '2008': 1}
I think the answer might be in: Most will remember September 2008, which was
when the credit crisis got really ugly as Lehman Brothers went bankrupt and
interest rates surged.
```

But the system doesn't do well with descriptive questions like:

```
Question:  What affects GDP?
{'affect': 1, 'affects': 1, 'GDP': 1}
I think the answer might be in: "The revision of 2013 GDP could affect the
size of 2014 GDP but will basically not affect GDP growth for 2014," the
bureau said in a statement.
```

Further work needs to be done to NER tag the CEO-Type questions with CEO tagger and Company-Type questions with Company tagger. Then the exact entities could be matched and extracted as the specific answers. For the descriptive questions, further processing needs to be done to identify the 'intent' of the question. Also, answering such questions might require processing multiple top sentences.

Methodology

Corpus Indexing:

The entire corpus was initially taken, and documents were extracted. The documents were indexed into ElasticSearch.

Lifecycle of a Question:

1. **Question Analysis:**
   First step is to identify the question type using a rule-based approach. Each question is associated with a list of nltk NER-tagger compatible tags.
   Keywords are extracted from the question by removing stopwords and adding in the lemmatized form of the other words.
   The documents matching these keywords are then matched and scored and top 10 documents returned. Since ElasticSearch provides this functionality out-of-the-box, we used a match query to extract top 10 documents from the index created in the previous step.
2. **Answer Extraction and Analysis:**
   Next step is to flatten the documents into a list of sentences.
   The sentences are NER tagged using nltk NER tagger and the ones which don't contain any of the tags identified in the previous step, are ignored.
   Once the candidate sentences are extracted, they are scored based on the following criteria:
   - The number of words in the candidate sentence that occur adjacently in both the question and the answer candidate (+)
   - The TF-IDF sum of the number of words that matched between question and answer (+)
   - The TF-IDF sum of the number of question content words that did not match in the answer (-)
   The top sentence after scoring is returned as the answer sentence.

Further Work:

Further work is based on specific domain of specific question types. For the question types specified the below could be further done:

- Which companies went bankrupt in month X of year Y?
  A Company NER tagger can be utilized to find lists of companies in say top 5-10 sentences and return a list of companies.

- Who is the CEO of company X?
  A CEO NER tagger can be utilized to extract the CEO name

- What affects GDP? What percentage of drop or increase is associated with this property?
  This question will require a lot of domain specific processing to be done on top of extracted sentences.

Analysis

The following results were obtained on a few test cases of each question types:

Who is the CEO of company X?

```
Question:  Who is the CEO of Google?
{'CEO': 1, 'Google': 1}
I think the answer might be in: Larry Page (CEO of Google) Net Worth: $32.3
billion 4.

Question:  Who is the CEO of Facebook?
{'CEO': 1, 'Facebook': 1}
I think the answer might be in: Mark Zuckerberg, CEO of Facebook, when
Facebook's IPO flopped, the talk was that the young Zuckerberg wasn't ready
to be a strong leader of a major company.

Question:  Who is the CEO of Twitter?
{'CEO': 1, 'Twitter': 1}
I think the answer might be in: In addition to being Twitter's CEO, Evan
Williams was its largest shareholder.
{'CEO': 1, 'Amazon': 1}
I think the answer might be in: Prime members can also borrow more than
700,000 books, listen to one million songs and hundreds of playlists, save
unlimited photos and watch tens of thousands of movies and TV episodes
including the Golden Globe nominated show from Amazon Studios, Transparent,
said Jeff Bezos, founder and CEO of Amazon.com.
```

Which companies went bankrupt in month X of year Y?

```
Question:  Which companies went bankrupt in month September of year 2008?
{'company': 1, 'companies': 1, 'go': 1, 'went': 1, 'bankrupt': 1, 'month': 1,
'September': 1, 'year': 1, '2008': 1}
I think the answer might be in: Most will remember September 2008, which was
when the credit crisis got really ugly as Lehman Brothers went bankrupt and
interest rates surged.

Question:  Which companies went bankrupt in month may of year 2009?
{'company': 1, 'companies': 1, 'go': 1, 'went': 1, 'bankrupt': 1, 'month': 1,
'may': 1, 'year': 1, '2009': 1}
I think the answer might be in: In 2009, it went bankrupt, shuttered all but
one of its 33 plants in the US, let go 25,000 workers, and shifted its center
of gravity to Shanghai.

Question:  Which companies went bankrupt in month december of year 2001?
{'company': 1, 'companies': 1, 'go': 1, 'went': 1, 'bankrupt': 1, 'month': 1,
'december': 1, 'year': 1, '2001': 1}
I think the answer might be in: In the early 2000s, the industry was in
turmoil, with a glut of fiber being built across the U.S. Before its collapse
amid an accounting scandal, Enron Corp dabbled in fiber routes, while
WorldCom was losing money on its fiber networks before its top executives
committed fraud and the company went bankrupt.
```

What affects GDP? What percentage of drop or increase is associated with this property?

```
Question:  What affects GDP?
{'affect': 1, 'affects': 1, 'GDP': 1}
I think the answer might be in: "The revision of 2013 GDP could affect the
size of 2014 GDP but will basically not affect GDP growth for 2014," the
bureau said in a statement.

Question:  What affects interest rates?
{'affect': 1, 'affects': 1, 'interest': 1, 'rate': 1, 'rates': 1}
I think the answer might be in: LSAPs, in contrast, most directly affect term
premiums...Â As both forward rate guidance and LSAPs affect longer-term
interest rates, the use of these tools allows monetary policy to be effective
even when short-term interest rates are close to zero The FOMC also has
attempted to credibly communicate to the market "the criteria that would
inform future decisions about the program."

Question:  What affects inflation index?
{'affect': 1, 'affects': 1, 'inflation': 1, 'index': 1}
I think the answer might be in: How Does the Value of the Dollar Influence
Inflation?One important way the dollarâ□□s value affects inflation is through
commodity prices.
```

Conclusions

The Question and Answer system works well for factual questions which have specific answers. It would be appropriate to apply domain knowledge to question types if we know the system will be asked only such questions. This will require specific NER taggers to be used over the extracted sentences to output the precise correct answers.

For the questions which require a more descriptive answer, the system needs to process much more sentences and extract specific information asked from the system.

Next Steps

The following issues need to be addressed as next steps:

1. Clean corpus before adding to ElasticSearch index and searching based on cleaned keywords.
2. Use CEO and Company NER taggers to extract CEO and Company names from the top sentences and output as the answer.
3. Write further processors for descriptive questions.