**Question-Answering system**

**Introduction and Objective:**

The objective of this project was to create a simple Question-Answer system that will answer the following 3 types of questions:

1. Who is the CEO of company X?
2. What companies were bankrupt in month X of year Y?
3. What affects GDP and by how much does this factor affect GDP?

Questions were posed and answers were retrieved from a collection of Business Insider articles over the years 2013 and 2014

**Methodology:**

The same general method was used to address all 3 question types as follows:

*Preprocessing:*

First a corpus of documents was created by reading all the articles into Python and saving them as a single corpus. The files were each converted into ASCII form before being read in. Furthermore, a dictionary was created for each file recording labels such as date created.

*Question analysis:*

Next after tokenization and preprocessing of questions such as deletion of stop words, a query was generated by extracting keywords from the question and determining the question type such as whether it asked for a name, date etc. Keywords were determined from part of speech tagging to extract types such as proper nouns, nouns, adjectives and in general capitalized words or quoted words. The following keywords were generated for each question:

Question 1: CEO, chief, Executive and capitalized words (names)

Question 2: Bankrupt, name of month X, name of month Y, Inc., co., company

Question 3: GDP, percentage, %

*Document analysis:*

The document was first preprocessed similarly to how the question was dealt with. Stop words removed, tokenization and tagging were all done in a similar fashion. A term frequency table was created on the documents that contained at least one of the keywords from the respective questions. Now after discarding the rest of the documents the selected ones were scored using the term frequency criterion where the frequency of each keyword was counted and recorded in a list for each document. Now this list of documents was recorded with the top scoring documents in terms of keyword frequency sorted first. In order to make the program faster, I eliminated the bottom 60% of ranked articles at this stage since they were most likely irrelevant to the question being posed. Now another table was created that contained inverse document frequency scores for each term. Again documents were ranked according to those with the highest IDF score depending on how exclusive the keywords were to each document. Essentially we create TF-IDF scores for each document.

*Sentence extraction, analysis and ranking*

From the list of the most relevant documents determined in the previous step, each documented was segmented into sentences using the sentence tokenizer. Now each sentence was analyzed as follows. First each was parsed to check if at least one of the keywords exists. If yes, the sentence was saved and if no, the sentence was discarded. From the selected sentences, the sentence type was determined using NER and only the ones that matched the type determined from question analysis were saved. This reduced the overall sentences to 5% of the total selected ones from the previous step. To now rank this list of selected sentences a method was used that combined TF-IDF and Cosine similarity. First the TF-IDF values were determined for each of the selected sentences in the same way it was done for the documents. Each sentence had a vector output of TF-IDF values. Now cosine similarity was applied to the vectors to determine which sentences most accurately matched the question type to give a good answer.

**Evaluation**

From the results of each of the questions we can see that in many cases the first few responses have included the most important keywords from the question and in some cases has answered the specific question type. However, despite there being the correct facts in the answer, on many times the answer is not related to the question and often contains a large and complex sentence that happened to contain all keywords. Additionally, precision and recall scores were not analyzed to check how accurate and inclusive the responses were. In order to improve this question answer system in the future that generates answers that are more specific and answer the question more concisely, more analysis would be done on the answer type using extensive NER. Here we will be able to focus on answers that directly answer the question in a specific concise form. Additionally, more analysis would be done on keyword generation. Many relevant answers would not necessarily contain all the keywords in the question but would contain instead related words. Identifying related words is a big challenge and requires more time. However, this would increase the recall rate in general.

**Results:**

***Type 1:***

***Who is the CEO of Tesla?***

*1. SolarCity, of which Tesla CEO Elon Musk is chairman, is also up 2.5%.*

*2. Elon Musk, CEO of Tesla, who disregarded the naysayers, paid off a huge department of energy loan early, and is generally regarded as one of the most inspirational people in business today.*

*3. AP ImagesClosely-followed hedge fund manager David Einhorn, the CEO of Greenlight Capital who is taking aim at Apple over its massive $137 billion cash hoard, just wrapped up a conference call with Apple shareholders.*

*4. The outspoken automotive executive, who made the comment on the network's "Squawk Alley" program, echoed a sentiment expressed by Tesla-founder Elon Musk earlier this month.*

*5.  Shares of Tesla jumped earlier this week on reports that CEO Elon Musk had met with Apple's Adrian Perica, an executive who does merger and acquisition deals for the iPhone-maker.*

***Who is the CEO of Facebook?***

*1.Mark Zuckerberg, the owner of Facebook, is not your typical corporate CEO.',*

*'"Facebook is basically a mobile company right now a majority of its revenues come from mobile, not from the desktop product," explained our Jim Edwards.*

*2.You can catch the release when it goes live here FREE AppDownload\r\nREUTERS/ Chip EastHedge fund titan Carl IcahnSome people think Facebook's $19 billion purchase of WhatsApp is insane.*

*3. Zuckerberg's desk and his laptop, because he's posted videos from there before*

*4.Peter Thiel, a founder of PayPal, an internet payment company, and the first outside investor in Facebook, a social network, says that innovation in America is "somewhere between dire straits and dead*

***Who is the CEO of Snapchat?***

1*. Snapchat CEO Evan Spiegel made candid comments about the tech sector and Facebook in particular in a Nov. 2013 email to a Sony executive who is a Snapchat board member, as revealed in the huge Sony hack.*

*2.Snapchat/Google PlayThe founders of a red-hot messaging company called Snapchat are suing each other over ownership of the company.*

*3.We called Sam Shank who runs HotelTonight and asked him what Evan Spiegel, Snapchat's CEO, might be thinking.*

*4. Take the case of Evan Spiegel, the 23-year-old CEO of a startup called Snapchat. One tech executive I met with last week shook his head in bewilderment when Snapchat came up*

***Type 2:***

***What companies went bankrupt in December 2013?***

*1. Bank Everbright defaulted on an interbank loan 10 days ago amid wild spikes in short-term "Shibor" borrowing rates, a sign that liquidity has suddenly dried up.*

*2. China Everbright Bank reportedly defaulted on 6 billion yuan of an interbank loan from Industrial Bank Co. earlier this month.*

*3. GT Advanced, slated to supply scratch-resistant sapphire displays from a newArizonafactory Appleis helping bankroll, filed for bankruptcy on Monday but has offered little explanation as to why.*

*4. After 20 months of thorough investigations by the Department of Justice, two bankruptcy trustees, and the CFTC, no evidence has been found that contradicts Mr. Corzines sworn testimony before Congress.*

***What companies went bankrupt in March 2014?***

*1.* *Gox said in its February 28 bankruptcy filing that hackers may have exploited what it called "a bug in the bitcoin system" to steal virtual currency from the exchange.*

*2. Gox filed for bankruptcy, a half dozen employees at the Tokyo-based bitcoin exchange challenged CEO Mark Karpeles over whether client money was being used to cover costs, according to three people who participated in the discussion.*

*3. TheCredit Suisse Group AG (CSGN)unit said it was confident it could arrange the proposed bankruptcy-exit loan as long as LightSquared, a wireless broadband provider, met conditions including obtaining the cooperation of all parties-at-interest in the reorganization and all required regulatory approvals.*

*4.Gox sought Chapter 11-style bankruptcy protection, saying it had lost bitcoins and cash worth some half a billion dollars due to hacker attacks.*

***What companies went bankrupt in October 2013****?*

*'The country sank 100bn into feed-in tariffs or in solar companies that blazed the trail, did us all a favour, and mostly went bankrupt, displaced by copy-cat competitors in China*

*2. The Credit Suisse Group AG (CSGN)unit said it was confident it could arrange the proposed bankruptcy-exit loan as long as LightSquared, a wireless broadband provider, met conditions including obtaining the cooperation of all parties-at-interest in the reorganization and all required regulatory approvals.*

*3. From the time all the trouble started in late 2012 to the summer of 2013 right before OGX went bankrupt Batista’s companies lost combined $9.7 billion*

*4. 'Some pundits have noted that LIBOR spiked before Lehman Brothers went bankrupt in 2008*

*Type 3:*

***What affects GDP?***

*1.* *The result was the strong increase in the ratio of total debt (ex financials) to GDP for emerging economies, by a staggering 36% since 2008.*

*2. I urge you not to roll your eyes at the attempt to understand yet another dry economic statistic, but instead to look deeply at how the attempt to measure GDP affects everything in our lives*

*3. In the U.S., the U.K., and other rich nations, the percentage of gross domestic product going to wages continues to decline while the percentage going to profits steadily increases*

*4. in the fourth quarter,current-dollar GDP increased 1.3 percent, or $53.1 billion*

*5. 'The GDP results for the first half are a more accurate indicator of how weak the economic recovery has become*

***How much does debt affect GDP?***

*1.* *Deutsche Bank's Joe LaVorgna is impressed: Significantly narrower November international trade deficit (-$34.3B vs. -$39.3B) pushes our estimate of Q4 real #GDP growth to +4.0%.*

*2. Morgan Stanley estimates that at $80 a barrel which the investment bank gives a 45% probability the government's budget deficit would increase to 2% of GDP next year, while inflation would continue to climb to 9%*

*3. 'The overall average result shows that public debt overhang episodes (over 90% GDP for five years or more) are associated with 1.2% lower growth as compared to growth when debt is under 90%*

*4. Correct: It now looks like GDP will increase in the 2.2% to 2.4% range for 2013.1)Question #1 for 2013: US Fiscal Policy [T]the House will fold their [early 2013] losing hand [on the debt ceiling] soon*

***How much do wages affect GDP?***

*1.* *In the U.S., the U.K., and other rich nations, the percentage of gross domestic product going to wages continues to decline while the percentage going to profits steadily increases.*

*2. In spite of a further large rise in the base this year, the GDP growth has subsided noticeably and corporate profits after taxes and adjusted for inventory gains/losses (IVA) and over/under depreciation (CCA) has declined 10% in the latest four quarters*

*3. Since labor reforms and falling wages have made hiring a more palatable proposition for employers, analysts said, GDP growth of even 0.5 percent might jumpstart employment growth*