Matthew Shipley

IEMS 308

Professor Klabjan

Homework Assignment 4

Question & Answer System

Summary

By using several concepts from text mining and analytics, a question and answer system was designed utilizing two years of news articles as a corpus from which to draw information. The question and answering system is able to handle four types of questions: asking which companies went bankrupt in month X of year Y, asking which factors affect the national GDP, asking what percentage change these factors are associated with, and asking who the CEO of well known companies are.

Functions and Methodology

Elasticsearch_dsl – Elasticsearch was used to index and search the corpus. Elasticsearch allows searches to be made over the entire corpus extremely quickly, so the Q&A system can be quite responsive.

extract_entities() – The extract_entities() function accepts a sentence input and returns any names and organizations discovered in the string. This function utilizes the nltk package to token, part-of-speech tag and chunk into entities.

classifyQuestion() – The classifyQuestion() function accepts a sentence input and returns which of the four question types the question is categorized into. This function uses a cosine similarity algorithm to score each of the four question types based on similarity to the candidate input, and assigns the question to one of the four categories based on which type scores the highest.

gdpPercent() – The gdpPercent() function answers questions of the percentage change in GDP type. This function first extracts the search term from the input question, and queries the Elasticsearch index for this target term and other terms like “GDP”, “percent”, “%”, “growth”, and others. Sentences are ranked by counting the presence of these terms, with some weighted more heavily (such as the search term) in order to create an accurate ranking. The top 3 candidate sentences based on score are chosen, and each word in those three sentences is then scored again based on traits like proximity to the word GDP and the target word and whether the word is “%” or “percent”. The highest scoring percent figure is returned.

companyBankrupt() – The companyBankrupt() function answers questions of the bankruptcy type. The function first extracts the month and year from the input question, and searches the Elasticsearch index for this date and bankruptcy keywords. Sentences are filtered based on the year and month that are stated in the sentence, and sentences are then scored based on inclusion of the target date and keywords like “bankruptcy”, “Chapter 11”, and “filed”. The top scoring sentences are compiled, and the extract_entities() function is used to compile the company names. The company name with the highest score is returned.

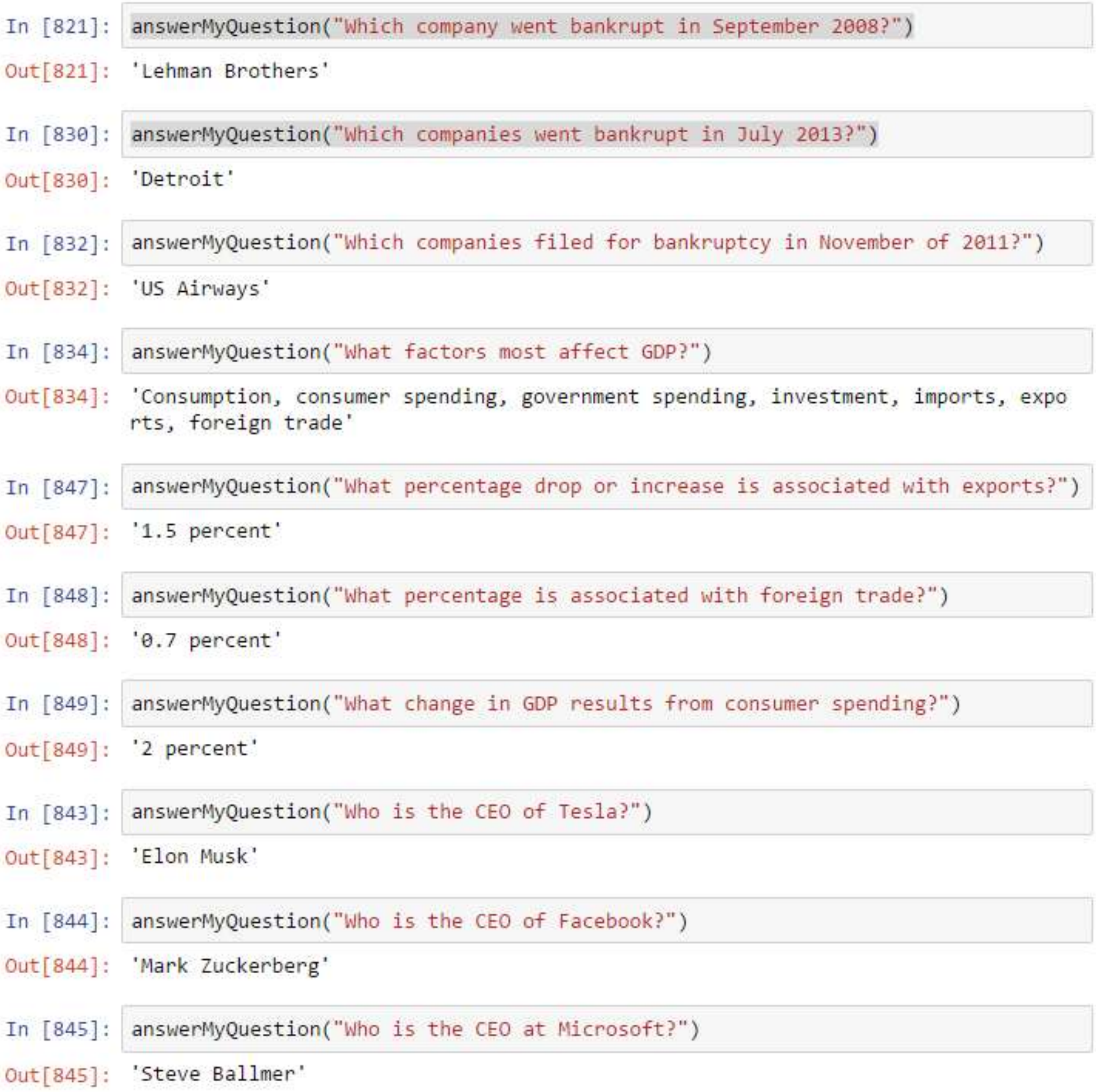
findCEO() – The findCEO() function answers questions of the CEO name type. The function first extracts the company name from the input question using the extract_entities() function. The corpus is then searched for the company name. Sentences are filtered out based on whether the word "CEO" is in the sentence. Then the extract_entities() function is used again to take out the names from the top scoring sentences, and names matching the company name itself are removed. The name that appears in the plurality of sentences including the word "CEO" and the company name is returned.

answerMyQuestion() – The answerMyQuestion() function accepts any question type and outputs the answer to the question. This function first uses the classifyQuestion() function on the input string, and then executes the appropriate function out of gdpPercent(), companyBankrupt(), and findCEO().

Instructions for Use

- Make sure Elasticsearch is installed by extracting the Elasticsearch .zip file, opening the command prompt, navigating to the /elasticsearch-5.0.2/bin folder, and typing "elasticsearch"
- Open ShipleyQASystem.ipynb in Jupyter Notebook
- Make sure to install the nltk and elasticsearch packages, as well as any packages not installed that are displayed in the first cell of the notebook
- In the third cell of the notebook, change the "location" variable to be equal to the location of your 2013 and 2014 folders that contain the news articles. The first initialization should be where the 2013 articles are, and the second should be where the 2014 articles are
- If Jupyer Notebook and Elasticsearch are running properly, then click Cells/Run All. This should execute each cell in the notebook, which initializes all function necessary to run the answerMyQuestion() function. The indexing of the corpus will take around 5 minutes to complete
- To run the Q&A system, create a new cell and run the answerMyQuestion(str) function, replacing str with your question in quotations, i.e. "Who is the CEO of Tesla?"
- The function should return Elon Musk!

Sample Questions



```
In [821]:  answerMyQuestion("Which company went bankrupt in September 2008?")

Out[821]:  'Lehman Brothers'

In [830]:  answerMyQuestion("Which companies went bankrupt in July 2013?")

Out[830]:  'Detroit'

In [832]:  answerMyQuestion("Which companies filed for bankruptcy in November of 2011?")

Out[832]:  'US Airways'

In [834]:  answerMyQuestion("What factors most affect GDP?")

Out[834]:  'Consumption, consumer spending, government spending, investment, imports, expo
           rts, foreign trade'

In [847]:  answerMyQuestion("What percentage drop or increase is associated with exports?")

Out[847]:  '1.5 percent'

In [848]:  answerMyQuestion("What percentage is associated with foreign trade?")

Out[848]:  '0.7 percent'

In [849]:  answerMyQuestion("What change in GDP results from consumer spending?")

Out[849]:  '2 percent'

In [843]:  answerMyQuestion("Who is the CEO of Tesla?")

Out[843]:  'Elon Musk'

In [844]:  answerMyQuestion("Who is the CEO of Facebook?")

Out[844]:  'Mark Zuckerberg'

In [845]:  answerMyQuestion("Who is the CEO at Microsoft?")

Out[845]:  'Steve Ballmer'
```