IEMS 395 Homework Assignment 4
Question / Answer System
Nick Paras

Executive Summary

The purpose of this assignment was to create a simple question and answer system to extract relevant sentences from a text document corpus based on a user's input query. The corpus in question is 2 years of postings on BusinessInsider.com. This corpus is comprised of 35298 individual postings of varying lengths. Although the methodology is generalized to accept any query (see methodology), the Q/A system is designed to be able to answer 4 specific types of queries: "Who is the CEO of *x*?", where *x* is a company; "Which companies went bankrupt in *y*?", where *y* is a Month Year pair such as September 2008; "What affects GDP?"; and "What percentage is associated with *z*?", where *z* is one of the factors that affects GDP.

The general approach is characterized by a series of actions that iteratively rank and reduce the search space of text data until an answer (or set of answers) is retrieved. First, a corpus is created out of the text documents, where each document contains a single post on BI.com. Then, the type (in this case, CEO, company, percentage, etc.) of query and the keywords in the query are extracted from the question posed by the user. A function performs 'document retrieval' using those keywords and selects only the documents with at least one of the keywords in the query. Then, a document-term matrix is created out of these remaining documents, and it is weighted according to Tf-Idf values (this implementation actually uses the "apc" variant of SMART Tf-Idf because it achieved better results) to score these documents based on their relevance to the query. Then, these documents are ranked by score, and the top 10 are chosen. Now, all sentences are extracted. From here, only the sentences that share at least one word with they keywords are selected. Next, the type of each sentence is determined. If a sentence is of a different type than the query from the original question, it is discarded. The remaining sentences are once again converted to a document term matrix and scored. This time each sentence is scored, and the top 10 are returned.

These top 10 sentences are the answer candidates, and can be presented several different ways. One could either return the full sentences, or one could perform Named Entity Recognition to extract just the company names, CEO names, etc. from those sentences. Upon the suggestion of Professor Diego Klabjan, this question/answer system uses the full sentences. This allows the user to see the context of each answer and choose the response that is most appropriate for their needs.

This approach is valuable because it demonstrates a generalized methodology for extracting information from unstructured text data. In a business or research context, this proves to be extremely valuable, because it allows a user to quickly find an answer to a question without having to read thousands of documents manually. Many of us take this capability for granted in our daily lives since the advent of popular search engines like Google.com or Bing.com, but this method allows for the creation of specialized engines built for specific corpora. A simple example might be a corpus of a particular business' service agreements and customer complaints. This type of question answer system would allow an engineer, manager, or any other employee to access information quickly, eliminating the expense of many man-hours of reading, while simultaneously improving internal process efficiency and potentially even customer relations and service.

Introduction

The purpose of this assignment was to create an automated Question/Answer system for a corpus of 2 years of Business Insider postings. This Question/Answer system should be able to answer 4 specific types of queries: "Who is the CEO of *x*?", where *x* is a company; "Which companies went bankrupt in *y*?", where *y* is a Month Year pair such as September 2008; "What affects GDP?"; and "What percentage is associated with *z*?", where *z* is one of the factors that affects GDP. This assignment implements the general question/answer system shown in the lecture slides QA-1.pdf using the *tm* package in R.

Data Exploration

The corpus is originally comprised of 730 text documents; one for each day spanning 2013 and 2014. These documents are included in the folder `./allPostings` and are encoded in CP1252. These are then split such that there is a separate document for each postings using the shell script `reformatPostings.sh` and the new text files are located in `./reformattedPostings`. There are now 35398 text files. Upon import into R and conversion into a corpus, 100 empty files are removed, resulting in a total of 35298 posts considered for this assignment.

Methodology

As mentioned above in Data Exploration, the original text files are split so that each document contains one and only one post. These documents are then used to create a corpus in R of size 219 Mb. A copy of this corpus is made and is then preprocessed (stemming, punctuation removal, etc.). At this point, 100 empty documents are identified by creating a document term matrix. Any document that has 0 occurrences for every token in the corpus is removed from both corpora and a new document term matrix is created. The corpora and the document term matrix are made global to the R workspace (so that subsequent functions may still access them).

Query Type and Keyword Extraction

The next step is to determine the type and keywords of the user's question. The assignment does not require that system can answer any generic question, but rather just the four types of factoid questions posed in the introduction. Therefore, a function is defined that uses a rule-based approach and regular expressions to determine the type of the question. This function is quite simple; for example if the word CEO appears, it is classified as a CEO type question. Then, another function is defined that extracts the keywords from the query. This function uses the Part of Speech tagging classifier from the *openNLP* package to identify the nouns, adjectives, and Proper nouns of the query for extraction. This is performed in *getKeywords()* and *getQueryType()*.

At this point it is worth noting that several additional keywords are added depending on the query type to augment the vector of keywords. This is done because the Q/A system is specialized for those four types of queries. However, it is worth noting that the Q/A system will still function for other queries, it will just skip the keyword augmentation. This is performed in *keyAug()*.

Document Retrieval and Ranking

The next step is to identify the documents that have at least one of the keywords and to exclude the rest. This is done by making a document-term matrix, identifying the tokens that match the keywords, and selecting the documents that have a (strictly) positive sum of occurrences of those

tokens. This is performed in *getDocuments()*. Then, these documents are used to create a new document-term matrix, which is weighted by its SMART Term Frequency-Inverse Document Frequency score. This is an augmented version of the traditional Tf-Idf score used in text-analytics, but adjusts the formula with a positive bias parameter, a probabilistic denominator rescaling, and a cosine normalization scheme. While not always a better or worse choice than traditional Tf-Idf scores, the SMART scheme performed better for this application. These weights are added for each term in the query for each document, and the top 10 documents are chosen. This is performed in *scoreDocuments()*.

Sentence Extraction and Scoring

Now I have extracted the top 10 most relevant posts (documents) according to the weighting function defined above. The next step is to break up each of these posts into individual sentences. This is performed in *getSentences()*. Now, a new corpus is made out of those sentences, and a new document term matrix is created and weighted (again using the SMART Tf-Idf discussed above). As before, sentences that don't contain any of the keywords are removed. This is performed in *getSentenceMatches()*.

Next, sentences are classified by type (CEO, company, etc.) using a mixture of NER and Regular Expressions. For example, the *openNLP* Maxent_Entity_Annotator NER function is used to identify sentences with percentages. It can also be used for Peoples' names and organization names extraction. However, since the model is trained (It is a maximum entropy/ logistic regression classifier) on English data, it does not work reliably on non-english names. For example, it cannot recognize Microsoft CEO Satya Nadella, or the company Tumblr. One more option would be to use the CEO and Company classifiers we trained last week, however they proved to be very time intensive and regular expression extractions to find adjacent capitalized words `("[A-Z]+[a-z]*\ [A-Z]+[a-z]+[A-Z]*[a-z]*")` actually performed better. Once the sentences are classified by type, only the ones that match the query time are selected. This is performed in *pruneSent()*.

After removing the extra sentences, a new corpus is made (and preprocessed) and the sentences are scored as shown in class. The scoring method is a weighted sum of the number of keyword matches plus the Tf-Idf (again, SMART variant) score of the matches minus the Tf-Idf score of the non-matches. The weights were tuned manually to 5, 2, and 1, respectively. This is performed in *scoreSentences()*.

Finally, the sentences are ranked by their scores and the top 10 are returned. At this point, one could use NER to extract the relevant names based on the type of the query, however after talking with Professor Klabjan, this system elects instead to return the whole sentences. This gives the user additional context to ensure they choose the best answer, rather than simply accepting the single top scoring answer.

Instructions for Use

First either unzip the canvas submission and navigate to that working directory, or clone the github repository with the clone URL https://github.com/ngparas/Question-Answer-System.git To load all of the functions and initialize the Q/A system, one should load the file `sourceCode.R`, and execute all lines up until the section labeled `# Sample Queries`. At this point, the question answer system is ready to execute. One can now ask a question by entering the command:

```
questionAnswer("Your Question Here")
```

into the R console. As noted above, this Question Answer system will accept any query, although it should be noted that the performance for questions of types other than those outlined in the introduction might vary considerably.

Results

The assignment asks us to provide 9 sample questions and answers. This part is included in the source code under the section labeled `# Sample Queries`:

```
# 3 sample CEO queries
questionAnswer("Who is the CEO of Apple?")
questionAnswer("Who is the CEO of Facebook?")
questionAnswer("Who it the CEO of Microsoft?")

# 3 sample bankruptcy queries
questionAnswer("Which companies went bankrupt in September 2008?")
questionAnswer("Which companies went bankrupt in October 2011?")
questionAnswer("Which companies went bankrupt in September 2014?")

# GDP Query
questionAnswer("What affects GDP?")

# 3 sample follow-up queries
questionAnswer("What percentage is associated with personal consumption?")
questionAnswer("What percentage is associated with exports?")
questionAnswer("What percentage is associated with federal defense spending?")
```

For brevity, the actual output of each question is not included here but rather included in **Appendix A**. But, for the sake of discussion, it is worth noting that the correct answer does appear at least once in the top 10 answer candidates for all 10 of these sample queries.

Discussion and Limitations

As can be seen in Results and Appendix A, the Q/A system appears to perform fairly well. It found the correct answer for all 9 of the sample test cases. However it is worth noting that there are a number of limitations. Some of these limitations are a result of the corpus and the question types chosen, and finally some are a result of the implementation.

<u>Corpus/Question Type Limitations</u>

The performance of the system on questions of the type "Which companies went bankrupt in Month Year" is the weakest out of all of the question types. Most likely, this is because of the format of the information in the corpus. By exploring the raw corpus using the unix command `grep` one can see that most often times when companies are mentioned as going bankrupt, it is in the format "Company x filed for bankruptcy Tuesday." This makes sense because a human reader knows the date the article was written and can then infer the relative temporal distance intuitively, however it makes it difficult for a keyword-search algorithm such as the one implemented here to find those sentences when half of the keywords (the Month and Year) do not appear in most instances in the corpus. To some extent augmenting the keyword "case", which helps to identify the sentences that contain the court filings, alleviates this issue. Further, extracting the year from the keywords and

reformatting it in the "08-13..." format used in the court filing serial number helps to lend additional context. This dramatically improved the results. However, it is still limited by the contents of the corpus. More generally, the performance of all question types was adversely affected by the prominence of typos in the corpus. This limits the effectiveness of NER and some samples can be seen in Appendix A.

Implementation Limitations

Further, even if there is an answer to a question, such as:

```
questionAnswer("Which companies went bankrupt in July 2013?")
```

and there is an answer:

```
American Roads, which owns and operates toll roads in the U.S. and Canada, filed for bankruptcy
in July 2013, in part because traffic volumes fell during the recession despite projections in
2006 that they would rise.
```

it may not be (and in this case is not) found. There are a number of possible reasons for this. One reason could be that the document that it is found in is very long, so the document may not be retrieved at all because of Tf-Idf normalization. Another reason could be because the sentence itself is relatively long, the negative term in the sentence scoring function (Tf-Idf weights of non-keywords) could be sufficiently large to move it out of the top 10.

This can result in answers that are not necessarily correct despite their high scores being returned in the top 10. This is precisely why I elect to return the whole sentences—because it allows the user to verify that an answer is indeed correct and what they are looking for before accepting it as a fact.

Finally, this model could potentially be greatly improved if the NER capabilities could be expanded. Currently, NER is used for type-tagging purposes, but only alongside additional regular expressions. Ideally, one would use the NER classifiers we built in assignment 3, however these are prohibitively computationally intensive to scale to a corpus of this size without using a server to pre-tag each sentence. Therefore, the *openNLP* packages and Regular Expressions are used instead.

Conclusion

The purpose of this assignment was to implement a Question/Answer system for 2 years of Business Insider postings to answer 4 kinds of factoid queries. To complete this task, I implemented the generalized question/answer approach described in class in R using the *tm* and *openNLP* packages. I was able to achieve relatively good results, as shown in Results and Appendix A, however one should note that there are limitations to this type of system that stem from both the underlying texts and the system itself. That being said, this sort of system offers a vast amount of business value if it can be implemented effectively.

This approach is valuable because it demonstrates a generalized methodology for extracting information from unstructured text data. Many of us take this capability for granted in our daily lives since the advent of popular search engines like Google.com or Bing.com, but this method allows for the creation of specialized engines built for specific corpora. A simple example might be a corpus of a particular business' service agreements and customer complaints. This type of question answer system would allow an engineer, manager, or any other employee to access information quickly, eliminating the expense of many man-hours of reading, while simultaneously improving internal process efficiency and potentially even customer relations and service.

Appendix A – Full Console Output for 9 Sample Questions

```
> # Sample Queries ------------------------------------------------------
>
> # 3 sample CEO queries
> questionAnswer("Who is the CEO of Apple?")

Query Type: person
Query Keywords: CEO Apple
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "APApple CEO Tim Cook announcing the new iPhones."
 [2] "Fossil shares were down more than 3%, while Movado shares were off 1.5%, after Apple announced Apple Watch."
 [3] "Is AppleCare a waste of money or a good way to protect your costly Apple products?"
 [4] "Is AppleCare a waste of money or a good way to protect your costly Apple products?"
 [5] "And Apple easily crushed expectations on the top and the bottom lines."
 [6] "Apple Pay has a chance at changing the mobile payments game completely."
 [7] "APApple CEO Tim Cook.Here's the funny thing about Apple's earnings report from last night: No matter what you thought of the company, the report can easily justify your opinion."
 [8] "If Apple hits the target of 80 million, it would be 33% year-over-year growth, which would be very strong for Apple's iPhone business, which has had growth in the single digits and low double digits lately."
 [9] "AP ImagesClosely-followed hedge fund manager David Einhorn, the CEO of Greenlight Capital, added to his Apple stake in the fourth quarter, according to a 13F filing."
[10] "GT Advanced Technologies, which makes sapphire displays, was down as much as 11% after Apple announced that sapphire displays would be in Apple Watch, but not the company's newest iPhones: iPhone 6 and iPhone 6 Plus."

> questionAnswer("Who is the CEO of Facebook?")

Query Type: person
Query Keywords: CEO Facebook
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Follow BI Video: On Facebook"
 [2] "TumblrFacebook CEO Mark ZuckerbergFacebook isn't just popular with its 1.3 billion (and counting) users."
 [3] "Facebook has a sales office in Hong Kong."
 [4] "Facebook's founder and CEO Mark Zuckerberg owns about 500 million shares of Facebook stock."
 [5] "Microsoft's CEO asked Mark Zuckerberg in 2007."
 [6] "Produced by Matthew Stuart Follow BI Video: On Facebook"
 [7] "Time Warner CEO Jeff Bewkes nixed the idea."
 [8] "Stephen Lam/ReutersFacebook CEO Mark ZuckerbergCan’t stand the big boss at work? Try getting a job at LinkedIn, Facebook, or Starbucks."
 [9] "Stephen Lam/ReutersFacebook CEO Mark ZuckerbergCan’t stand the big boss at work? Try getting a job at LinkedIn, Facebook, or Starbucks."
[10] "During the Spring of 2005, Facebook (still TheFacebook) was talking to The Washington Post Company about an investment."
```

```
> questionAnswer("Who it the CEO of Microsoft?")

Query Type: person
Query Keywords: CEO Microsoft
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Business Insider/Julie BortMicrosoft CEO Steve BallmerMicrosoft earnings are out!"
 [2] "APMicrosoft CEO Satya Nadella.Microsoft reported its Q2 2014 earnings Tuesday afternoon."
 [3] "CEO Satya NadellaMicrosoft earnings are out, and they're solid."
 [4] "Microsoft FREE AppDownload"
 [5] "Microsoft just announced CEO Steve Ballmer will retire."
 [6] "We recently came out with our list of the Sexiest CEOs Alive!"
 [7] "We recently came out with our list of the Sexiest CEOs Alive!"
 [8] "In August, Microsoft CEO Steve Ballmer announced he will be stepping down after 13 years
leading the company."
 [9] "Then a few weeks later CEO Steve Ballmer retired."
[10] "We spoke with Chris Suh, head of investor relations at Microsoft after the report."
>
> # 3 sample bankruptcy queries
> questionAnswer("Which companies went bankrupt in September 2008?")

Query Type: organization
Query Keywords: companies bankrupt September 2008
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Lehman Brothers collapsed in September 2008 in the highest-profile failure of a bank during
the financial crisis."
 [2] "Lehman Brothers Holdings Inc, the Wall Street bank that filed for bankruptcy on September
15, 2008, on Thursday is returning $17.9 billion to creditors, boosting its payout so far to
$80.4 billion in five distributions."
 [3] "The case is In Re: Trump Entertainment Resorts Inc, U.S. Bankruptcy Court, District of
Delaware, No:14-12103."
 [4] "\"The Trustee brings this action against Defendants for acts and omissions that culminated
in the business collapse of the Company and the bankruptcies of the Debtors."
 [5] "The cases are In re: MF Global Inc, U.S. Bankruptcy Court, Southern District of New York,
No. 11-02790; and In re: MF Global Holdings Ltd in the same court, No. 11-15059."
 [6] "Romney opposed an auto bailout in a November 2008 op-ed in the New York Times, which was
memorably titled, \"Let Detroit Go Bankrupt.\""
 [7] "HJ Heinz, the son of immigrant parents, built this this company from scratch."
 [8] "By Jonathan Stempel NEW YORK (Reuters) - Former customers of MF Global Holdings Ltd's
bankrupt brokerage will recoup all $6.7 billion they are owed following the completion of a
payout that will begin on Friday, its trustee said."
 [9] "APEarlier, today Warren Buffett's Berkshire Hathaway announced it and 3G Capital would
acquire HJ Heinz Company in a transaction worth a whopping $28 billion."
[10] "Here's the chart: Gallup The score reflects 37% of workers telling Gallup that their
employer is hiring, and 15% saying their company is letting people go and reducing the size of
its workforce."

> questionAnswer("Which companies went bankrupt in October 2011?")
```

```
Query Type: organization
Query Keywords: companies bankrupt October 2011
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

[1] "Market worries about that exposure was among the factors that led to MF Global's quick collapse and October 31, 2011, bankruptcy."
[2] "The case is In Re: Trump Entertainment Resorts Inc, U.S. Bankruptcy Court, District of Delaware, No:14-12103."
[3] "\"The Trustee brings this action against Defendants for acts and omissions that culminated in the business collapse of the Company and the bankruptcies of the Debtors."
[4] "Broker-dealer MF Global collapsed in the fall of 2011 and more than a billion in customer funds went missing."
[5] "The Internal Revenue Service told Delphi in June that it would be taxed as a U.S. company due to the sale of its assets to Delphi Holdings LLC after it emerged from bankruptcy in 2009, the company said in a regulatory filing on July 31."
[6] "HJ Heinz, the son of immigrant parents, built this this company from scratch."
[7] "APEarlier, today Warren Buffett's Berkshire Hathaway announced it and 3G Capital would acquire HJ Heinz Company in a transaction worth a whopping $28 billion."
[8] "\"We will continue to prepare and file our financial statements on the basis that neither Delphi Automotive LLP nor Delphi Automotive Plc is a domestic corporation for U.S. federal income tax purposes,\" the company said in the filing."

> questionAnswer("Which companies went bankrupt in September 2014?")

Query Type: organization
Query Keywords: companies bankrupt September 2014
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

[1] "The case is In Re: Trump Entertainment Resorts Inc, U.S. Bankruptcy Court, District of Delaware, No:14-12103."
[2] "\"The Trustee brings this action against Defendants for acts and omissions that culminated in the business collapse of the Company and the bankruptcies of the Debtors."
[3] "REUTERS/Joshua LottModels relax before presenting creations from the Michael Kors Spring/Summer 2014 collection during New York Fashion Week, September 11, 2013.Good morning."
[4] "HJ Heinz, the son of immigrant parents, built this this company from scratch."
[5] "These forecasts are based on the July SPCS data release this morning and the August 2014 Zillow Home Value Index (ZHVI), released September 18. Officially, the SPCS Composite Home Price Indices for August will not be released until Tuesday, October 28."
[6] "S&amp;amp;P Dow Jones Indices managing director David Blitzer said, \"After a long period when home prices rose, but at a slower pace with each passing month, we are seeing hints that prices could end 2014 on a strong note and accelerate into 2015.\""
[7] "APEarlier, today Warren Buffett's Berkshire Hathaway announced it and 3G Capital would acquire HJ Heinz Company in a transaction worth a whopping $28 billion."
[8] "REUTERS/Jason Redmond Members of a medical marijuana delivery service promote their business at the High Times U.S. Cannabis Cup in Seattle, Washington September 8, 2013."
>
> # GDP Query
> questionAnswer("What affects GDP?")

Query Type: gdp
```

```
Query Keywords: GDP
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Personal consumption grew by 3.0%, adding 2 percentage points of growth to GDP."
 [2] "Meanwhile exports fell 7.6% taking a full percentage point of growth from GDP."
 [3] "Here's a breakdown of the components of GDP: BEA Here's how much the various components of
GDP added to growth."
 [4] "The biggest factors weighing on growth were nonfarm private inventories, which subtracted
1.70 percentage points from GDP growth, and federal defense spending, which subtracted 1.28
percentage points."
 [5] "Each component of GDP varies widely in how much it contributes to growth."
 [6] "Arenamontanus / FlickrThe European Union now expects GDP growth in the euro-area to fall
0.3 percent in 2013."
 [7] "YouTubeQ4 GDP unexpectedly fell 0.1 percent."
 [8] "GDP rose 0.1 percent, missing economists' predictions of a bigger 0.5 percent gain."
 [9] "U.S. GDP grew at a measly 0.1% rate in Q1."
[10] "The biggest positive drivers were consumption of durable goods, which contributed 1.01
percentage points to GDP growth, and nonresidential fixed investment in equipment and software,
which contributed 0.79 percentage points."
>
> # 3 sample follow-up queries
> questionAnswer("What percentage is associated with personal consumption?")

Query Type: percentage
Query Keywords: percentage personal consumption
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Personal consumption grew by 3.0%, adding 2 percentage points of growth to GDP."
 [2] "This was largely driven by personal consumption growth, which was revised up to 2.0% from
1.4%."
 [3] "The personal saving rate – personal saving as a percentage of disposable personal income –
climbed to 3.2% from 3.0% in April."
 [4] "Core personal consumption expenditures gained 1.5% in June, a tick ahead of estimates for
1.4%, and 0.1% month-over-month, down from the 0.2% rate in June."
 [5] "Expectations were for personal spending to rise by 0.1 and for personal income to rise by
0.3%."
 [6] "Personal consumption growth came in at 2.1 percent, below both the advance estimate (2.2
percent) and expectations (2.3 percent)."
 [7] "Core personal consumption expenditure prices came in as expected, unchanged at 0.1% growth
month-over-month, and up 1.1% year-over-year versus 1.2% last month."
 [8] "Expectations were for personal income to rise 0.3% and personal spending to rise 0.4%
month-over-month."
 [9] "Last month, personal spending rose by 0.5% and personal income rose by 0.3%."
[10] "Last month, personal spending unexpectedly declined 0.1%."

> questionAnswer("What percentage is associated with exports?")

Query Type: percentage
Query Keywords: percentage exports
Retrieving Documents
```

```
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Meanwhile exports fell 7.6% taking a full percentage point of growth from GDP."
 [2] "Export growth of China and Taiwan was only 4.9% and 4.4% yoy in 2Q14 respectively, so a
boost of 1 to 2 percentage points to headline export growth is no small matter, especially for
the currency market which closely tracks export growth numbers.\""
 [3] "Export to the US fell 0.8 percent."
 [4] "Personal consumption grew by 3.0%, adding 2 percentage points of growth to GDP."
 [5] "Exports to the EU fell 11.1 percent."
 [6] "The employment sub-index however surged to 55.8, from 52.5, up 3.3 percentage points."
 [7] "Exports to China fell 15.8 percent."
 [8] "Exports fell by 5.8 percent year-over-year versus the estimate of -4.2 percent."
 [9] "iPhone 6 sales are also expected to boost Taiwan's export growth 2% per month from August
through October, and 1% from November through January."
[10] "The first shows that of the top 20 export markets for the U.S. eleven are developing
nations \"but only two—Brazil (ranked 7th) and India (18th)—are of any significance when it comes
to the current problems in the emerging markets.\""

> questionAnswer("What percentage is associated with federal defense spending?")

Query Type: percentage
Query Keywords: percentage federal defense spending
Retrieving Documents
Scoring Documents with SMART TF-IDF
Retrieving Sentences
Pruning Sentences by Type
Scoring Sentences
The 10 best scoring answers (in descending order) are:

 [1] "Compensation of nondefense employees and civilian defense employees makes up about one-
fifth of real federal spending and about 1.5% of GDP."
 [2] "Federal spending on nondefense items was actually up 1.4 percent."
 [3] "The biggest factors weighing on growth were nonfarm private inventories, which subtracted
1.70 percentage points from GDP growth, and federal defense spending, which subtracted 1.28
percentage points."
 [4] "Federal spending fell 15.0 percent, led by a 22.2 percent drop in defense spending."
 [5] "State and local spending fell 0.7 percent."
 [6] "National defense decreased 0.7 percent, compared with a decrease of 0.6 percent."
 [7] "Government spending was the largest driver of the economic contraction in the fourth
quarter, subtracting 1.33 percentage points from Q4 GDP growth and falling 6.6 percent."
 [8] "The investment surge added 1.18 percentage points to Q4 GDP growth."
 [9] "Net trade subtracted 0.81 percentage points from Q4 GDP growth."
[10] "National defense decreased 22.2 percent, in contrast to an increase of 12.9 percent."
```