Executive Summary

The purpose of this project is to build a Q&A system that will extract specific information from a corpus built out of articles from Business Insider over 2013-2014. There are four main questions that the program is equipped with answering. “Which companies went bankrupt in month X of year Y? What affects GDP? What percentage drop or increase in GDP is associated with this property? Who is the CEO of company X?”.

On a high level, a rule-based approach is employed to answer these questions. A combination of Okapi score and tf-idf scores are used to extract the appropriate documents from the corpus and appropriate sentences from the documents. The classifier that was built from Assignment 3 was also used to further cut down the number of potentially valid answers. Depending on the question type, some customization to the answer selection algorithm is made. For example, questions like “Which companies went bankrupt in month X of year Y”, we had to enforce a strict filter to only keep sentences that contains the word “bankrupt” or “bankruptcy.

In general, the quality of the results that we were able to gather differs depending on the question type. “Who” questions tend to outperform “Which” and what questions. This might be because sentences containing CEO names does not usually contain other name-like words, making it easier to pinpoint the answer. And sentence related to companies and GDP contain a lot of conflicting information that would often produce confusing results. A more detailed rationale would be provided in more details later.

Methodology

1. Given a question, a certain set of features are extracted:
    a. The WH_word (What, Which, Who etc.),
    b. Keywords that are nouns, adjectives, and numbers.
2. Depending on the question, an extended list of related words are appended to the keywords. For example, words like “LLC, Holdings, Group etc.” is appended to keywords for company questions.
3. The next step would be to select a certain number of documents based on the Okapi score computed from the keywords extracted from the last step.
4. We then sentence tokenize the documents and use the classifier built from Assignment 3 to extract appropriate sentences-answer pairs. For the Who question, we want to only extract CEO names. For the Which question, we want to extract company names. For the What question, we want to extract percentages.
5. For most questions, a score is then computed using the formula (# of matched bigrams in sentence and keywords + sum of tf-idf of keywords in sentences) and the top answers are selected from sentence-answer pairs that produced a high score. We also initially wanted to subtract tf-idf scores of non keywords from sentences, but that did not work produce good results. For the “What affects GDP” question, we extracted answers by summing up the tf-idf scores of all words. Since factors that affects GDP tend to be

longer words, we also filtered out words that are short and common which might lead to high tf-idf scores.

6. Depending on the question. Some tailoring to the answer selection algorithm is made to help filter to the most appropriate results.
    a. For the What question, we have to filter to sentences that contain both “GDP” and the factor that affects GDP.
    b. For the Which question, we have to filter to sentences that contain “bankrupt” and “bankruptcy”.
7. Finally, sentences with the top 10 highest scores is displayed as potential answers to the question.

Results

Sample Outputs:

Q1. Who is the CEO of company X

Company Name	Tesla	Google	Facebook
Answer 1	**Elon Musk**	**Larry Page**	Facebbok Facebook Facebook Ahead
Answer 2	Tesla Motors	Google CEO Larry	**Facebook CEO Mark Zuckerberg**
Answer 3	Tesla CEO Elon Musk	Alan Mulally	Microsoft CEO Steve Ballmer
Answer 4	Andrew Mason	Salar Kamangar	Mark Zuckerberg
Answer 5	Nick Macfie	Michael Barrett	Mark Zuckerberg
Answer 6	Scott Olson	Dick Costolo	Mark Zuckerberg
Answer 7	Alex Richardson	Henrique De Castro	Facebook COO Owwn Van Natta
Answer 8	Larry Ellison	Larry PageGoogle	Jim Breyer
Answer 9	Bill Gates	Ross Levinsohn	Mark Zuckerberg
Answer 10	Kevin Ryan	Evan Williams	Jim Scheinman

Q2. What affects GDP?

- Government, Investment, Percentage, Appdownload, Inequality, Consumption, Contraction, Construction, Washington, Opportunity

Q3. What percentage of drop or increase in GDP is associated with X?

Company Name	Inequality	Consumption	Investment
Answer 1	17-40%	70%	1.18 percentage points
Answer 2	9.97 percent	2 percent	One percentage point
Answer 3	A percent	25 percent	A percentage
Answer 4	50%	14.8 percent	1.5 percent
Answer 5	A percentage	1.5 percent	14.8 percent
Answer 6	76%	1.9%	4%
Answer 7	43%	71%	23 percent
Answer 8	0.6%	As percentage	2 percent
Answer 9	1%	As percentage	56%
Answer 10	74%	5%	As percentage

Q4. Which companies went bankrupt in month August of Year 2013?

Company Name	March, 2015	June, 2014	August, 2013
Answer 1	Lehman Brother Holdings INc	LightSquared	Delphi Holdings LLC
Answer 2	NFL	ManhattanWASHINGTON	Delphi
Answer 3	Trustee	Harbinger Capital Partners	San Bernadino
Answer 4	MF Global holdings Ltds	LightSquared	Lehman Brother Inc
Answer 5	Stockton	Light Squared	US Bankruptcy Court
Answer 6	Lehman Brother Inc	Free AppDownload	Stockton
Answer 7	US bankruptcy court	Lehman Brothers Holdings Inc	FDIC

Answer 8	FDIC	NFL	Studios
Answer 9	Energy Future Holdings	Trustee	PwC
Answer 10	Reuters The TXU Monticello Steam Electric Station	MF Global Holdings Ltds	US Bankruptcy Court

Explanation to limitations in quality of results

The results for finding CEO names of companies are fairly accurate. However, the other questions did not see good results. This might be due to the following reasons:

1. In general, “Who is the CEO of company X” returns an answer with a singular correct answer that is easier to determine if it is right or wrong. However, for the other questions, the answer they are looking for are lies in a larger range. “What companies went bankrupt …” looks for a list of companies, “What affects GDP” looks for a list of factors, and “What percentage of drop…” looks for a range of percentage. As such, the larger range makes it harder for the algorithm to pinpoint one correct answer.
2. For “What percentage of drop or increase in GDP is associated with factor X”, the inaccuracy in results lies in the algorithms inability to discern whether the percentage is actually the effect on GDP associated with X. For example, a sentence can be “Unemployment dropped by 1.5%, GDP drop by 2%”. In this case, 1.5%, and 2% are all suitable candidates as an answer, but none of them actually answers the question.
3. For “Which companies went bankrupt in month X Year Y”. There are several difficulties that the algorithm found hard to overcome;
    a. The classification program I used for identifying companies seems to be not well-developed. Therefore, a lot of names that came up in the final results are not actual company names.
    b. The time frame associated with the question can not be effectively enforced in the document and sentence selection process. A lot of documents contain multiple years, leading to biased okapi score, which might lead to incorrect selection of documents. As a result, sentences that mentioned bankruptcy among the documents that were selected might not be actual bankruptcies that happened in the specified time frame.
    c. In a sentence containing the right time frame, and is related to bankruptcy, there might be several company names. The algorithm is unable to discern which company the bankruptcy is actually related to.

Conclusion / Business Insights and Value

Overall, what this project revealed is that the Q&A system we built utilizing Okapi scores for document selection and tf-idf scores for sentence selection is well equipped to answer

questions where there are very specific answers. However, the algorithm is less effective in seeking answers for more general questions that looks for a range of answers.

This means that the algorithm that we created would be of value for businesses that are looking to build a Q&A system that is dedicated to question types that returns specific answers rather than a range of answers. A more advanced Q&A system would need to be developed to help answer questions that returns more general answers. For example, the Q&A system might want to take into account not just the sentence alone, but the context surrounding the sentence. Better data extraction and processing can also be employed to better label the corpus according to their per-determined type. For instance, if a date-label could be extracted for documents in the original documents, we could use it to filter down to only documents in the relevant date range for “Which Companies went bankrupt in XXX”.