**Q&A**
**Zhiyuan Ping**

**Executive Summary**
There are 730 scraped articles for year 2013 and 2014. This corpus is used to extract all possible answers to the list of questions posted.

I first read in the 730 text files and tokenized the words while normalizing capital letters, punctuations and removing stop words. Then I built a hash map for the word and document number pairs, calculated the average document length and recorded the numbers representing the frequency of the most common words in a document. This completes the corpus pre-processing.

Since there are not training sets to build the cosine similarity algorithm for assigning types to questions, I simply identify the types of questions by searching for key words. Then I extracted the keywords for my Q&A system by normalizing, removing stop words, and identifying nouns followed by adjectives and below tags:
which = ["NNP","NNPS","RB","CD"]
what = ["NN","NNP","NNS","NNPS","VB","VBZ","VBD","VBN"]
who = ["NN","NNP","NNS","NNPS","VBD","VBZ"]

Then an algorithm that runs Okapi is performed to find the most relevant documents. Okapi is performed instead of tfidf because the documents have varied lengths. Then the most relevant documents are segmented in to sentences that represent the potential answers. An algorithm that scores sentences based on the sum of their Okapi scores and the number of common bigrams. Different parameters are attempted and all the results are uploaded; the final answers are in the “final_anwers.csv”.

**Questions**
Which companies went bankrupt in month September of year 2008?
Which companies went bankrupt in year 2009?
Which companies went bankrupt in year 2010?
What affects GDP?
What affects GDP? What percentage of drop or increase is associated with unemployment?
What affects GDP?
What affects GDP? What percentage of drop or increase is associated with interest rate?
What affects GDP?
What affects GDP? What percentage of drop or increase is associated with government spending?
Who is the CEO of Actavis?
Who is the CEO of Apple?
Who is the CEO of BlackRock?

**Q&A Interface**
The list of questions are stored in a csv file. The python code can be run directly to import the questions and perform the corresponding algorithm to find potential answers to the questions from the preprocessed corpus. The answers are saved to “final_anwers.csv”

**Data exploration and feature understanding**
The data in this project are the txt files, I turned it into a list of tokenized sentences and words. Features were extracted from the tokenized words and sentences based on their natures and relations to the neighboring tokens.

**Business Question**
The goal is to extract potential answers to a list of questions from a given corpus.

**Problem solution**
For each document or sentence, a score is calculated as bellows. Then the top ones are selected as candidates.

$$Okapi\ score = \sum_{t \epsilon Q,D} \left( ln \frac{N - df + 0.5}{df + 0.5} \right) \left( \frac{(k_1 + 1)tf}{\left( k_1 (1 - b) + b \frac{dl}{avdl} \right) + tf} \right) \left( \frac{(k_3 + 1)qtf}{k_3 + qtf} \right)$$

N = total number of documents in the collection

df = number of documents that contain the word

dl = document length

avdl = average document length

$k_1$ , $k_3$ , b are parameters

**Performance**

```
    ....
Which companies went bankrupt in month September of year 2008?
[('Which', 'JJ'), ('companies', 'NNS'), ('went', 'VBD'), ('bankrupt', 'JJ'), ('month', 'NN'),
('September', 'NNP'), ('year', 'NN'), ('2008', 'CD')]
This is a which question
The key words are:
['2008', 'september', 'bankrupt']
The docs selected are: 0     570
1      30
2     326
3     632
4     460
5     632
6     460
7     632
8     690
9     685
10    695
11    190
Name: 0, dtype: int64
There are 147 candidate sentences
The sentences selected are: [502, 503, 501, 3027, 5046, 7065, 2016, 2003, 2448, 2464]
The top sentence is CODA went bankrupt.
----------------------
Which companies went bankrupt in year 2009?
[('Which', 'JJ'), ('companies', 'NNS'), ('went', 'VBD'), ('bankrupt', 'JJ'), ('year', 'NN'),
('2009', 'CD')]
This is a which question
The key words are:
['bankrupt', '2009']
The docs selected are: 0     574
1     236
2      30
3      90
4     225
5      90
6     690
7      90
8     460
9     694
10    719
11    447
Name: 1, dtype: int64
There are 64 candidate sentences
The sentences selected are: [1744, 1746, 3119, 3106, 986, 3117, 1740, 3123, 1747, 1754]
The top sentence is No, France is not bankrupt ....
----------------------
```

```
----------------------
Which companies went bankrupt in year 2010?
[('Which', 'JJ'), ('companies', 'NNS'), ('went', 'VBD'), ('bankrupt', 'JJ'), ('year', 'NN'),
('2010', 'CD')]
This is a which question
The key words are:
['bankrupt', '2010']
The docs selected are: 0     386
1      74
2     236
3     147
4     271
5     147
6     325
7     147
8     477
9     644
10    356
11    607
Name: 2, dtype: int64
There are 56 candidate sentences
The sentences selected are: [2300, 1506, 2302, 1515, 1493, 2296, 11338, 2303, 2310, 734]
The top sentence is No, France is not bankrupt ....
----------------------
What affects GDP?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP')]
This is a what question
The key words are:
['gdp', 'affects']
The docs selected are: 0      55
1     367
2      74
3     325
4     477
5     325
6      69
7     325
8     442
9     684
10    190
11    251
Name: 3, dtype: int64
There are 77 candidate sentences
The sentences selected are: [6338, 2324, 3608, 5277, 7079, 2282, 7396, 8432, 4324, 1788]
The top sentence is The virus affects pigs but does not spread to humans.
```

```
What affects GDP? What percentage of drop or increase is associated with unemployment?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP'), ('What', 'WP'), ('percentage', 'NN'), ('drop',
'NN'), ('increase', 'NN'), ('associated', 'VBN'), ('unemployment', 'NN')]
This is a what question
The key words are:
['gdp', 'drop', 'unemployment', 'increase', 'percentage', 'affects', 'associated']
The docs selected are: 0     448
1     198
2     367
3     234
4     299
5     234
6     225
7     234
8     385
9     692
10    251
11    213
Name: 4, dtype: int64
There are 286 candidate sentences
The sentences selected are: [8560, 2194, 6004, 3112, 4478, 10217, 9268, 9478, 8530, 3884]
The top sentence is They then compared YLL over time with "the percentage change in daily mortality
associated with changes in air pollutants."
----------------------
What affects GDP?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP')]
This is a what question
The key words are:
['gdp', 'affects']
The docs selected are: 0     464
1     111
2     111
3     495
4     685
5     495
6     147
7     495
8     520
9     675
10    247
11    181
Name: 5, dtype: int64
There are 85 candidate sentences
The sentences selected are: [5207, 6164, 4204, 6023, 7571, 7781, 4845, 4507, 10164, 9870]
The top sentence is (3) Monetary policy affects both demand and supply.
----------------------
```

```
What affects GDP? What percentage of drop or increase is associated with interest rate?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP'), ('What', 'WP'), ('percentage', 'NN'), ('drop',
'NN'), ('increase', 'NN'), ('associated', 'VBN'), ('interest', 'NN'), ('rate', 'NN')]
This is a what question
The key words are:
['gdp', 'rate', 'interest', 'drop', 'increase', 'percentage', 'affects', 'associated']
The docs selected are: 0     145
1      26
2     292
3     385
4     148
5     385
6     264
7     385
8     147
9     686
10    240
11    708
Name: 6, dtype: int64
There are 553 candidate sentences
The sentences selected are: [2748, 5182, 7336, 2864, 5298, 7452, 8460, 10817, 3425, 8008]
The top sentence is They then compared YLL over time with "the percentage change in daily mortality
associated with changes in air pollutants."
----------------------
What affects GDP?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP')]
This is a what question
The key words are:
['gdp', 'affects']
The docs selected are: 0      77
1     524
2      26
3     599
4     442
5     599
6     396
7     599
8     556
9     413
10    519
11    713
Name: 7, dtype: int64
There are 77 candidate sentences
The sentences selected are: [9327, 2680, 4390, 5630, 3116, 4622, 6148, 4152, 1977, 465]
The top sentence is FREE AppDownload
Morgan StanleyIf oil prices stay as low as they are now, they will have profound affects on the
global economy.
```

```
----------------------
What affects GDP? What percentage of drop or increase is associated with government spending?
[('What', 'WP'), ('affects', 'VBZ'), ('GDP', 'NNP'), ('What', 'WP'), ('percentage', 'NN'), ('drop',
'NN'), ('increase', 'NN'), ('associated', 'JJ'), ('government', 'NN'), ('spending', 'NN')]
This is a what question
The key words are:
['gdp', 'drop', 'spending', 'government', 'increase', 'percentage', 'affects', 'associated']
The docs selected are: 0     457
1      84
2     524
3     229
4     147
5     229
6     477
7     229
8     225
9     680
10    332
11    248
Name: 8, dtype: int64
There are 379 candidate sentences
The sentences selected are: [3570, 6520, 6537, 3118, 2064, 5522, 1760, 2103, 2969, 4242]
The top sentence is Each index is calculated by subtracting the percentage of respondents reporting
a decrease from the percentage reporting an increase.
----------------------
Who is the CEO of Actavis?
[('Who', 'WP'), ('CEO', 'NNP'), ('Actavis', 'NNP')]
This is a who question
The key words are:
['ceo', 'actavis']
The docs selected are: 0     573
1     599
2      84
3      64
4      64
5      64
6      81
7      64
8     401
9     674
10    246
11     40
Name: 9, dtype: int64
There are 105 candidate sentences
The sentences selected are: [12262, 12259, 12708, 972, 999, 3023, 4892, 6761, 9129, 52]
The top sentence is Allergan had been holding off on talks with Actavis because it was waiting for a
California judge to decide on its request for an injunction to stop William Ackman, head of Pershing
Square, from voting at the shareholder meeting, a second source familiar with the situation
previously told Reuters.
```

```
----------------------
Who is the CEO of Apple?
[('Who', 'WP'), ('CEO', 'VBZ'), ('Apple', 'NNP')]
This is a who question
The key words are:
['ceo', 'apple']
The docs selected are: 0      50
1     641
2     599
3     664
4     637
5     664
6     331
7     664
8     374
9     646
10     35
11    579
Name: 10, dtype: int64
There are 259 candidate sentences
The sentences selected are: [10261, 9397, 9718, 10255, 10262, 10368, 10272, 10348, 10256, 9736]
The top sentence is He thinks that Apple's CEO is doing a great job.
----------------------
Who is the CEO of BlackRock?
[('Who', 'WP'), ('CEO', 'VBD'), ('BlackRock', 'NNP')]
This is a who question
The key words are:
['ceo', 'blackrock']
The docs selected are: 0     599
1       7
2       7
3     442
4     401
5     442
6     271
7     442
8     396
9     406
10    677
11     26
Name: 11, dtype: int64
There are 85 candidate sentences
The sentences selected are: [462, 1748, 5134, 9396, 461, 1747, 466, 1752, 12247, 5112]
The top sentence is From BlackRock: Conventional wisdom holds that more information is better.
```

**Analysis**

| | | | |
|---|---|---|---|
| 1 | CODA went bankrupt. | Fisker went bankrupt. | Better Place went bankrupt. |
| 2 | France is not and will not bankrupt because it would then be in a state of insolvency. | Now if France is bankrupt with a 0.3% GDP growth, what is the US with 0.1% shrinkage? | But a poll yesterday by Le Figaro newspaper showed eight out of ten readers agreed that France was indeed bankrupt. |
| 3 | Businesses would go bankrupt. | Here's the short version of what's happening: Cyprus's banks, like many banks in Europe, are bankrupt. | As such, banks don't want to lend, and these companies don't want to borrow for fear of going bankrupt before they even had a shot. |
| 4 | While severe winter weather often affects our third- | Rising inequality could also hurt the economy,as Wall | Real GDP has grown in 16 of 17 quarters, and the |

| | | | |
|---|---|---|---|
| | quarter results, the impact from multiple severe storms and frigid temperatures was significantly more pronounced this year and we are reducing our full-year earnings per share guidance as a result of the weather impact, warned CFO Alan Graf, Jr. | Street is starting to notice.As noted bySocit Gnrale strategist Albert Edwards, "you don't have to bea communist to conclude that high levels of inequality not only adversely affects long-term growth, but also increases the economys vulnerability to recession." | level of real GDP in the third quarter of 2013 was 5-1/2 percent above its pre-recession peak. |
| 5 | They then compared YLL over time with "the percentage change in daily mortality associated with changes in air pollutants." | With a savings rate of 8% (roughly that of the American economy) and GDP growth of 2%, wealth should rise to 400% of annual output, for example, while a drop in long-run growth to 1% would push up expected wealth to 800% of GDP. | And the only way we could have experienced a drop in the unemployment rate in the past year from 8.2% to 7.4% in the face of sub-2% economic growth strongly suggests that we are now on our way to a potential non-inflationary speed limit of just 1%. |
| 6 | This savings crisis affects individuals, families and the entire American economy. | They also mention lack of innovation, which affects brand desirability and ultimately investor sentiment and growth prospects. | Draghi also raised GDP projections in 2013 to -0.4% (up from -0.6%), and lowered 2014 GDP projections to 1.0% from 1.1%. |
| 7 | They then compared YLL over time with "the percentage change in daily mortality associated with changes in air pollutants." | Each index is calculated by subtracting the percentage of respondents reporting a decrease from the percentage reporting an increase. | The change in real private inventories added 0.59 percentage point to the second-quarter change in real GDP, after adding 0.93 percentage point to the first-quarter change. |
| 8 | And unemployment is at the heart of the macro dynamics that shape short- and medium-term inflation, meaning it also affects central banks. | "At our firm, the kind of event we're keeping an eye out for is a market correction that affects stocks that look like bonds--telecom stocks for example, that have high dividend yields and relatively low volatility," he writes. | In the upcoming year, the company's CFO, Mark Hood, said, "consumer confidence continues to be adversely impacted by ongoing macroeconomic headwinds, including health care costs and unemployment which disproportionately affects |

| | | | |
|---|---|---|---|
| | | | lower- and middle- income consumers." |
| 9 | Japan GDP beat estimates on an unexpected surge in capital spending. | This is the big risk  margin compression affects the E, while inflation, insofar as the tight historical relationship with final prices holds, even if to a smaller degree this time around, affects the P/E. | Here's his statement: Despite the 1% drop in real GDP in the first quarter, we believe that the US economy is now growing at an above-trend pace. |
| 10 | Actavis' CEO Brent Saunders will lead the company. | Actavis' CEO Brent Saunders addressed this argument head on in a CNBC interview after the Actavis-Allergan merger was announced. | The combined company will be led by Brent Saunders, CEO and President of Actavis, and Paul Bisaro will remain Executive Chairman of the Board. |
| 11 | Activist investor Carl Icahnissued an open letter to Apple CEO Tim Cook, in which he reiterated his call that Apple should buyback more stock and argued that Apple shares should be worth $203. | Leon Cooperman also spoke on Apple, as he was on live as Carl Icahn's open letter to Apple CEO Tim Cook crossed the wires, with Cooperman saying that Apple shares were currently 20% undervalued. | Here's the full letter: Dear Tim, As a large Apple shareholder with approximately 53 million shares, we applaud you and the rest of management, especially in light of recent launches and announcements which further validate our view that you are the ideal CEO for Apple. |
| 12 | BlackRock's Landers shares the same sentiment and sees opportunity in the selling pressure. | "Unfortunately, like a lot of conventional wisdom, its wrong," writes BlackRock's Russ Koesterich. | Dubbed "The guide that fund managers would love to ban", BestInvest names Scottish Widows, BlackRock, Baillie Gifford, F&C Investments and Jupiter as the big investment management houses that are failing to deliver. |

**Conclusion**

After comparing the results from running the algorithms on different parameters, I realized that the algorithm will achieve different accuracy for different type of questions. Answers to simpler questions such as bankruptcy and CEOs are easier to be found, compared to those to more

complex questions such as the effect of certain factor on GDP. The harder questions require more documents to be selected for sentence extractions. At the same time, Okapi is better performing than tfidf because it takes into account the impact of length on the calculated scores. For some reason, it is hard for the process to place emphasize on the text representing years. It can be a result of the corpus having more company names than years. Therefore the documents with company names but without years will play a dominant role in the selection process.

**Business insights and next steps**

With the algorithm developed, it is reasonable to assume that the plausible sentence selected will be at least relevant to the question, given the answers extracted after different trials with the different parameters. But the parameters need to be further fine-tuned for the algorithm to provide candidate sentence that exactly answer the questions, The parameters also have to be tailored for the specific question. Therefore, the next steps would be to build a more sophisticated algorithm that can target the question more specifically and identify the most relevant answer sentence in the corpus.

The business insights and value would be that if a corpus is known to contain certain topics or information, the algorithm can be trusted to produce the most relevant sentence to the question, providing an efficient way to extract information too certain topics of interests.