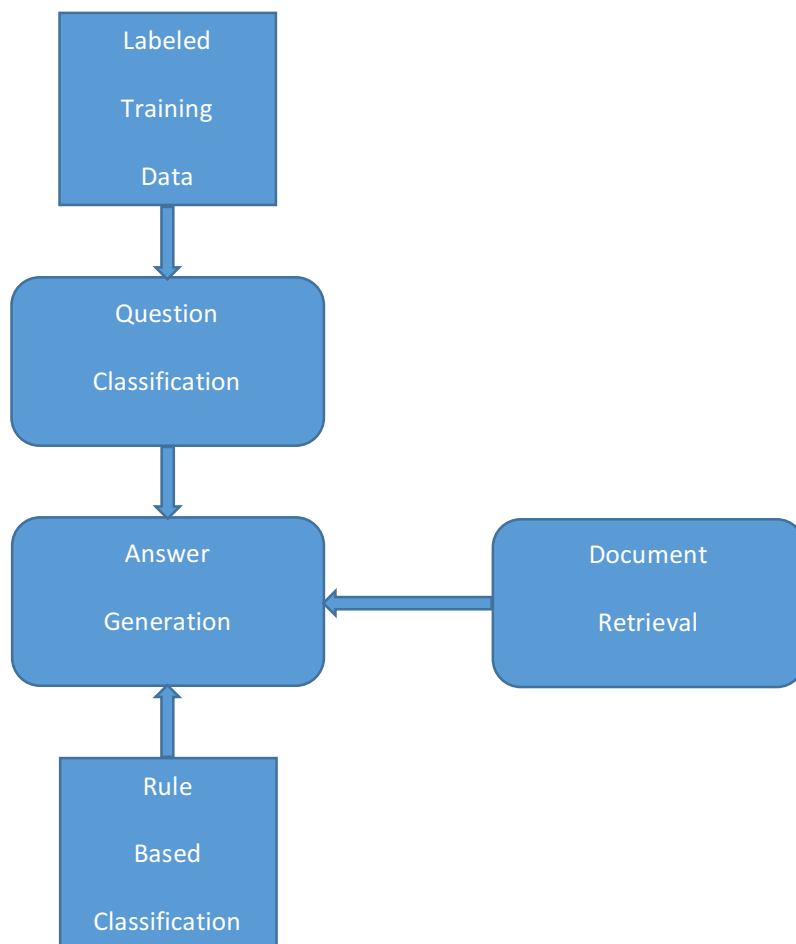# Hw4 Report

Zhaocheng Yu zyr531

Purpose:

Build a naïve question answer system to process couple of questions. Based on Business insider 2013 and 2014 daily news corpus, I build a simple question answer system from scratch. Python packages used include *nltk*, *sklearn* and etc. If your question is about CEO and company. You are welcome to type your question in line 166 at *Query_generation.py.* And run *answer_construction.py* to check the results. If your question is about GDP, you can still type your question in *Query_generation.py*. And then check all possible answer at *AnswerGeneration.py*.

Pipeline:

```
┌──────────────┐
│   Labeled    │
│   Training   │
│     Data     │
└──────┬───────┘
       │
       ▼
┌──────────────┐
│   Question   │
│Classification│
└──────┬───────┘
       │
       ▼
┌──────────────┐      ┌──────────────┐
│    Answer    │◄─────│   Document   │
│  Generation  │      │  Retrieval   │
└──────▲───────┘      └──────────────┘
       │
┌──────┴───────┐
│     Rule     │
│    Based     │
│Classification│
└──────────────┘
```

Labeled Training Data and Question Classification:

From the **Experiment Data for Question Classification** by Xin Li and Dan Roth, I get couple of labeled question data sets. There are basic six classes that a question could belongs to. **ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC** and **OTHER**. Features are types of Wh_words, POS tagger of first and second words preceded by Wh_words and length of the question. Supervised learning algorithm is **Decision Tree**. All details could be found in **QuestionClassifier.py**. Accuracy and confusion table attached.

```
Accruacy:
0.676999266324
Confustion Table:
          precision     recall  f1-score    support

     0        0.45        0.46      0.46          84
     1        0.67        0.66      0.67        1177
     2        0.70        0.50      0.59        1755
     3        0.75        0.77      0.76        1194
     4        0.54        0.75      0.63         602
     5        0.70        0.98      0.82         640

avg / total  0.68        0.68      0.67        5452
```
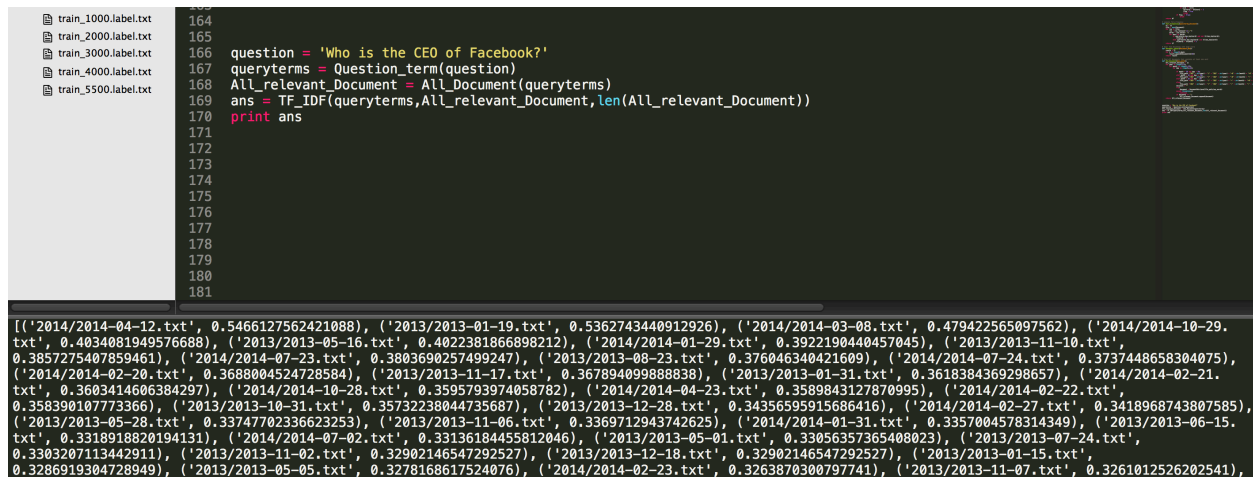
Figure 1 Question Classifier

Document Retrieval

Corpus used here are news of Business Insider from 2013-01-01 to 2014-12-31. Main approach here is to set up **question key words,** pick up documents that has at least one key words and **sort** those documents by **TF-IDF**. The example question is "Who is the CEO of Facebook? ''. Keywords are CEO and Facebook. The result returns all file names with corresponding TF-IDF

scores in descending order. You can find all details in **Query_generation.py.** Result of question "Who is the CEO of Facebook?" attached.


```
164
165
166  question = 'Who is the CEO of Facebook?'
167  queryterms = Question_term(question)
168  All_relevant_Document = All_Document(queryterms)
169  ans = TF_IDF(queryterms,All_relevant_Document,len(All_relevant_Document))
170  print ans
171
172
173
174
175
176
177
178
179
180
181
```

[('2014/2014-04-12.txt', 0.5466127562421088), ('2013/2013-01-19.txt', 0.5362743440912926), ('2014/2014-03-08.txt', 0.479422565097562), ('2014/2014-10-29.txt', 0.4034081949576688), ('2013/2013-05-16.txt', 0.4022381866898212), ('2014/2014-01-29.txt', 0.3922190440457045), ('2013/2013-11-10.txt', 0.3857275407859461), ('2014/2014-07-23.txt', 0.3803690257499247), ('2013/2013-08-23.txt', 0.376046340421609), ('2014/2014-07-24.txt', 0.3737448658304075), ('2014/2014-02-20.txt', 0.3688004524728584), ('2013/2013-11-17.txt', 0.367894099888838), ('2013/2013-01-31.txt', 0.3618384369298657), ('2014/2014-02-21.txt', 0.3603414606384297), ('2014/2014-10-28.txt', 0.3595793974058782), ('2014/2014-04-23.txt', 0.3589843127870995), ('2014/2014-02-22.txt', 0.358390107773366), ('2013/2013-10-31.txt', 0.35732238044735687), ('2013/2013-12-28.txt', 0.34356595915686416), ('2014/2014-02-27.txt', 0.3418968743807585), ('2013/2013-05-28.txt', 0.33747702336623253), ('2013/2013-11-06.txt', 0.3369712943742625), ('2014/2014-01-31.txt', 0.3357004578314349), ('2013/2013-06-15.txt', 0.3318918820194131), ('2014/2014-07-02.txt', 0.33136184455812046), ('2013/2013-05-01.txt', 0.330563573654083023), ('2013/2013-07-24.txt', 0.3303207113442911), ('2013/2013-11-02.txt', 0.32902146547292527), ('2013/2013-12-18.txt', 0.32902146547292527), ('2013/2013-01-15.txt', 0.3286919304728949), ('2013/2013-05-05.txt', 0.3278168617524076), ('2014/2014-02-23.txt', 0.3263870300797741), ('2013/2013-11-07.txt', 0.3261012526202541),

Figure 2 Candidates of Documents

Answer Generation and Rule Based Classification:

Due to the lack of labeled answer classification training data. I decided to use rule based method to classify answer. Sentence which has the NER tagger as '**PERSON**' can be selected to be the candidates of question whose label is '**HUM**' and etc. Besides, all answer candidates which obey the rule are still sorted by *TF-IDF* in descending way. You are welcomed to check all details in *AnswerGeneration.py*. Candidates of answer to 'Who is the CEO of Facebook?' attached.

['This chart also from Maude shows that revenue per user is also down on desktop: Ian Maude Revenue: Facebook Here\'s the ad revenue split over a longer timeframe (from BI Intelligence): Business Insider Intelligence Here are the monthly active users: Facebook The mobile monthly active users: Facebook And the mobile-only people: Facebook And here\'s how those users split from Business Insider Intelligence: Business Insider Intelligence Here\'s Zuckerberg\'s formal statement on the numbers: "Facebook\'s business is strong and growing and this quarter was a great start to 2014" said Mark Zuckerberg Facebook founder and CEO.', 'Here is the overall revenue breakdown: Facebook Here is a breakdown of Facebook\'s ad revenue per BII: Business Insider Intelligence Here\'s the overall user growth: FacebookAnd here\'s the user growth breakdown again from BII: Business Insider Intelligence "It was a great end to the year for Facebook" said CEO Mark Zuckerberg.', 'And Facebook CEO Mark Zuckerberg is doing exactly the right thing by investing heavily now instead of "maximizing profit" quarter after quarter.', "As it turns out CEO Mark Zuckerberg had other plans for continuing Facebook's outstanding growth that didn't require more ads and based on some recent data it appears those plans are working even better than expected.", 'The new "slow growth" era of Facebook will be compounded by the fact that both CEO Mark Zuckerberg and COO Sheryl Sandberg said that autoplay video ads and ads on Instagram \x97\xa0both sales products that analysts had presumed would add hundreds of millions in new revenue \x97\xa0won\'t be happening any time soon.', 'Follow BI Video: On Facebook', 'Facebook Creative Lab apps are just getting started.', ' Facebook posted good fourth-quarter results yesterday.', 'Other than that Facebook is entering an era of tough comparables.', 'Well it seems Facebook and I are in agreement!)', 'Two years ago before Facebook went public this is exactly what Facebook was—and the stock valuation soared as a result.', "It wasn't accusing Facebook of being the source of the fake clicks—anyone including Facebook might have created the bot clicks—but the allegation is that Facebook charged the advertiser for the fakes and it should have screened them out.", 'Perhaps Facebook is weeding out fake profiles and responding to my numbers after all.', "But unless Facebook CFO David Ebersman was just sandbagging everyone Facebook's profit

Figure 3 Candidates of questions

Results:

For question like who is the CEO of XX? and Which company went bankrupt in 20XX. We are able to get a relatively good result.

```
/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Who is the CEO of Facebook?
Mark Zuckerberg
[Finished in 20.7s]
```

```
/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Who is the CEO of Oracle?
Larry Ellison
[Finished in 19.6s]
```

```
/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Who is the CEO of Oracle?
Larry Ellison
[Finished in 19.6s]
```

Figure 4 Questions about CEO

```
/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Which company went bankrupt in 2013?
Lehman
[Finished in 23.3s]
```

```
/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Which company went bankrupt in August 2013?
Bitcoin
[Finished in 25.9s]
```

/Library/Python/2.7/site-packages/sklearn/cross_validation.py:44: DeprecationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from
that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
/Library/Python/2.7/site-packages/sklearn/utils/validation.py:395: DeprecationWarning: Passing 1d arrays as data is deprecated in 0.17 and will raise
ValueError in 0.19. Reshape your data either using X.reshape(-1, 1) if your data has a single feature or X.reshape(1, -1) if it contains a single sample.
  DeprecationWarning)
Which company went bankrupt in July 2013?
Reuters
[Finished in 26.1s]

Figure 5 Questions about bankrupt

['What happens in Connecticut affects only purchases here."', 'The grind size of your coffee beans: Grind size affects the extraction rate because it
affects surface area.', 'Temperature also affects flavor because it determines which solids get dissolved.', 'As representatives of \x93Main Street\x94
Congress cares about how Fed policy affects the average American.', 'The problem we have identified is not limited to MtGox and affects all transactions
where Bitcoins are being sent to a third party.', '"This is the big risk \xe2\x80\x93 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."', 'They also mention lack of innovation
which affects brand desirability and ultimately investor sentiment and growth prospects.', 'The temperature of your water: Temperature affects extraction
rate because solids dissolve more quickly at higher temperatures.', "There's also additional political uncertainty coming from other places like Ukraine
places like Argentina places like Thailand that affects markets.", '"Phonetic symbolism affects price perceptions because consumers typically process encode
and retain numbers (and hence prices) in memory in multiple formats" the authors write.', '"The peace and stability of this country has an impact on the
security of western China and more importantly it affects the tranquility and development of the entire region" Wang told a news conference alongside his
Afghan counterpart Zarar Ahmad Osmani.', 'In Antonopolous\' words: "As [Bitcoin] transactions are being created malformed/parallel transactions are also
being created so as to create a fog of confusion over the entire network which then affects almost every single implementation out there."', "It was noted
that in addition to the standard channels through which monetary policy affects the economy asset purchases could help signal the Committee's commitment to
accommodative monetary policy thereby making the forward guidance about the federal funds rate more effective." "Bernanke lays this dual strategy of

Figure 6 Answer Candidate of GDP

{'rate': 707, 'unemployment': 600}
['The unemployment rate had moved down in recent months as had broader measures of unemployment and underemployment.', 'The unemployment rate has been
steadily dropping for some time now.', 'The unemployment rate however remains elevated.', 'The unemployment rate had declined but remained elevated.', "Th
unemployment rate however remained elevated when judged against members' estimates of the longer-run normal rate of unemployment.", 'The unemployment
threshold is robust to this uncertainty.', 'With a trend decline in labor force participation and unemployment falling faster than expected Congress will
interested in just how far down the unemployment rate can sink.', 'In theory that also would push the unemployment rate lower.', "Gallup's unemployment ra
fell substantially between Feb and March.", 'Moreover potential changes to government unemployment benefits could drive unexpected shifts in the
unemployment and participation rates.']
[Finished in 25.0s]

Figure 7 Candidates of Unemployment