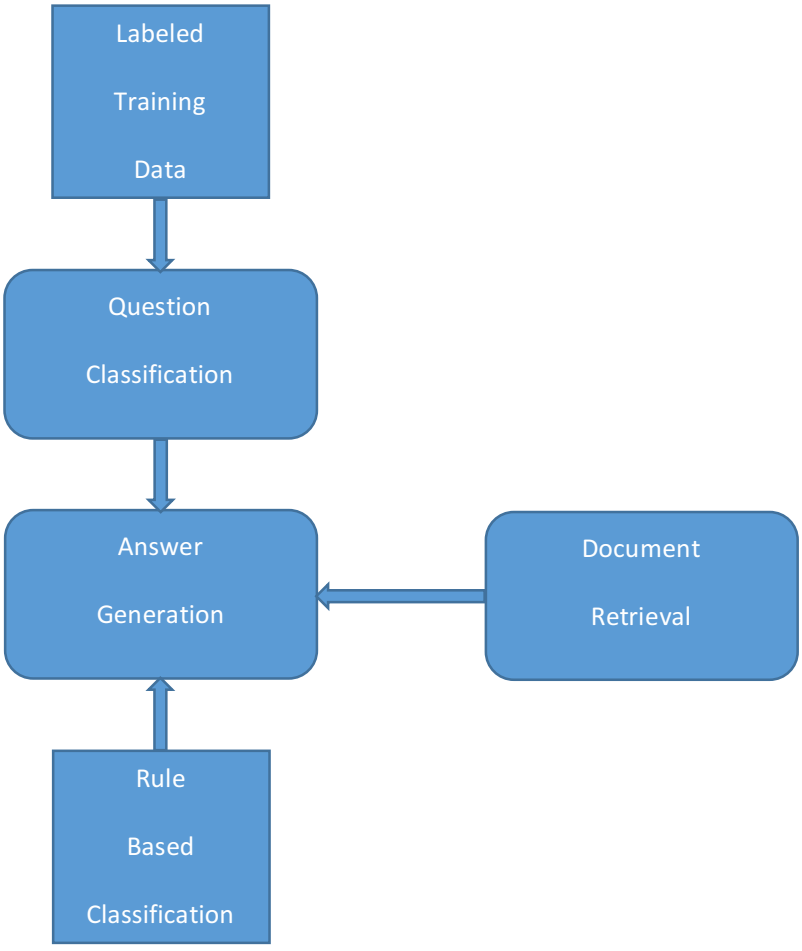
Hw4 Report

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***. It is because I failed to find a good way to deal with such a open question in this question answer system.

Pipeline:



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.

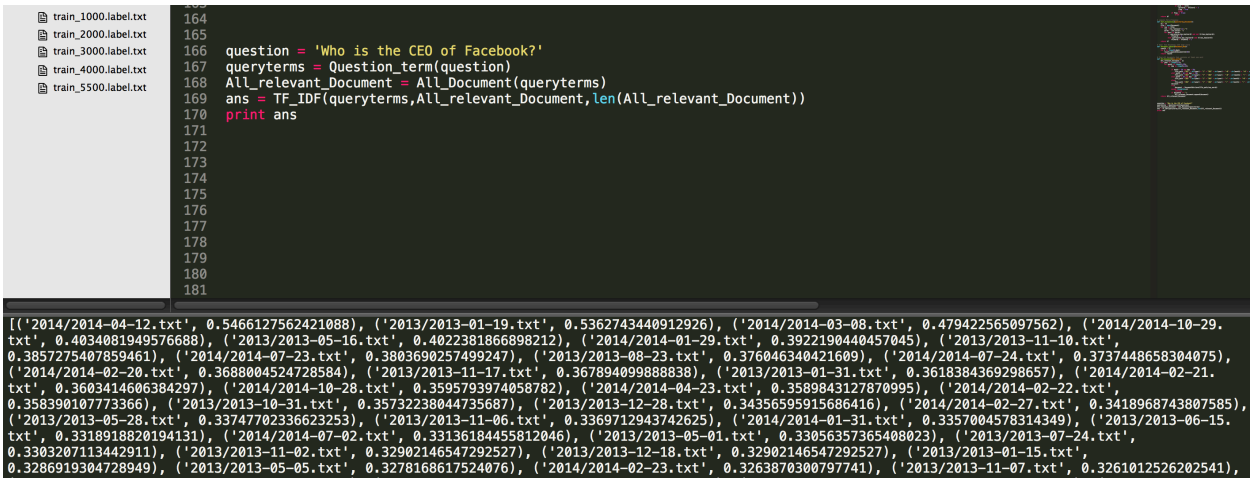


Figure 2 Candidates of Documents

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.

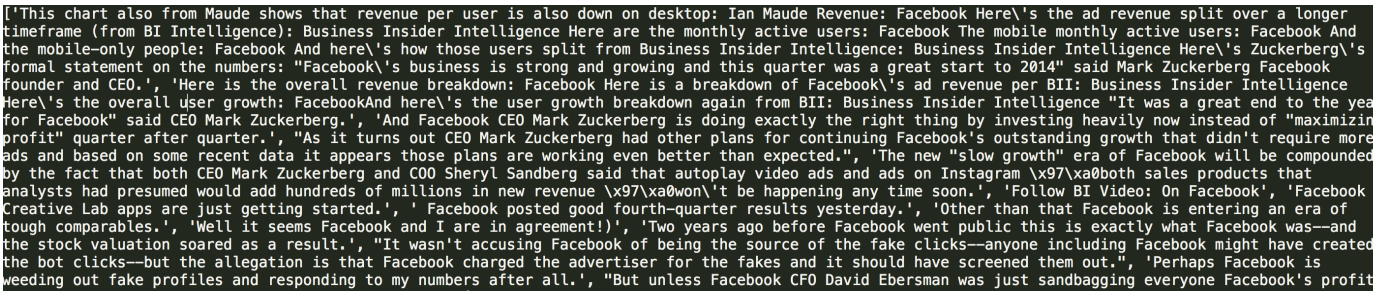


Figure 3 Candidates of questions

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]
```

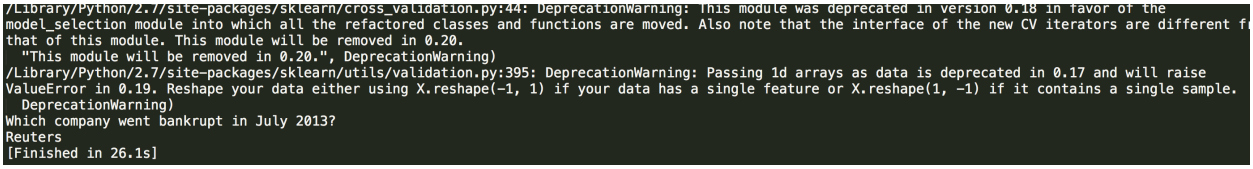


Figure 5 Questions about bankrupt

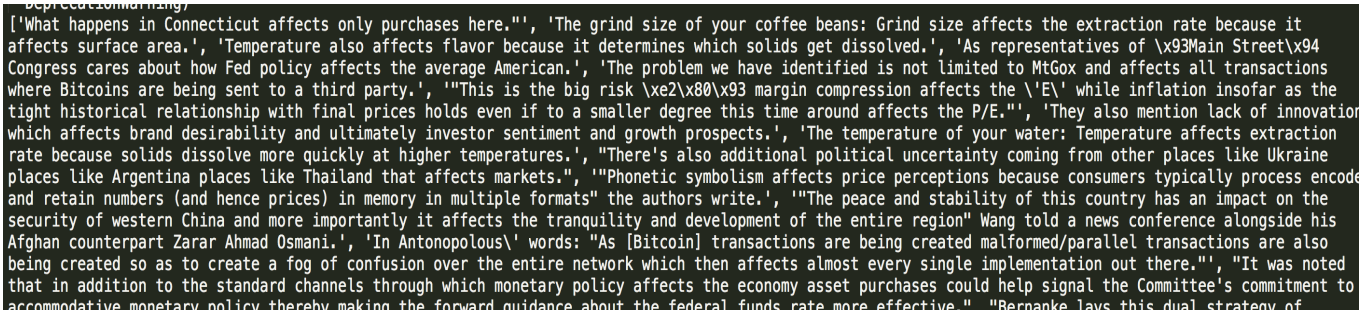


Figure 6 Answer Candidate of GDP

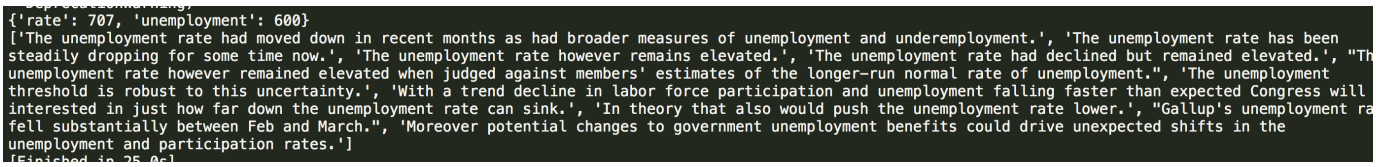


Figure 7 Candidates of Unemployment