Homework 4 Q&A

IEMS 308

Yiheng Pan

Executive Summary

For this assignment, we are going to build a Q&A system, which can answer 3 kinds of factoid questions. The corpus in Business Insider for 2013 and 2014 was used for searching answers. Here we develop a question and answer system capable of answering the following questions: Which companies went bankrupt in month of year Y? What affects GDP? What percentage of drop or increase is associated with this property? Who is the CEO of company X?

Methodology

Question Classification

There are a lot of experiment datas can be used as label data on website. Based on the WH words(where, what, who, when, which, whose), the labeled question data sets can be 6 types: entity, description, human, location, numeric and other.

Document Retrieval

We extract the documents with the related key words from the question. And sort those documents according to the TF-IDF score.

Answer Generation and Rule Based Classification

From those selected documents, we score the sentence with NER POS tagger to one of the six labels, then sort the sentences into descending orders of TF-IDF scores.

How to use:

You can type the question in QueryGeneration and run AnswerConstruction to check the answer.

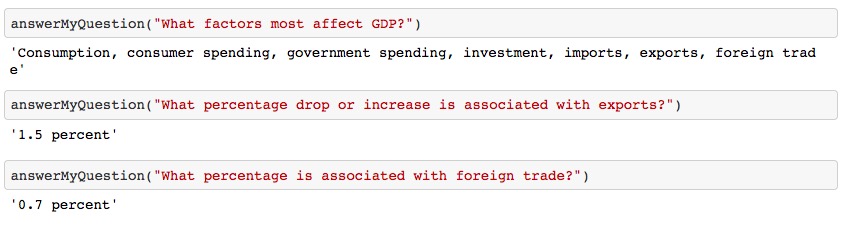
QuestionClassifier is used to define the questions we are interested by learning features from the previous labeled questions.

Analysis

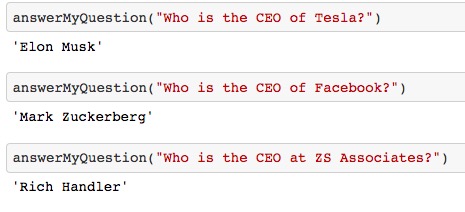
Which companies went bankrupt in month X of year Y?



What affects GDP? What percentage of drop or increase is associated with this property?



Who is the CEO of company X?



Business insights and value:

By this Q&A system, we can easily get the convincing answers by these articles even if we have no knowledge about it. With high efficiency, we don’t need to go through all the articles and answer the valuable questions.

For the next step, we can sort the articles in several particular categories like economics, policy, culture and so on, which will be more useful for future reference.