Jay Jung

IEMS 308

2016. 12. 08

Homework 4

Executive Summary

The purpose of this project was to create a simple Question-Answer system with the capability to answer 3 types of factoid questions. A collection of articles from Business Insider for the years 2013 and 2014 was used as a corpus. The basic search algorithm for the method was to parse through the corpus of articles to find documents that contain any of the keywords. Subset of documents were selected as the scope of search. Sentences that did not match the given question type in those subsets of data were disregarded. Sentences with matching key words were then scored by tf-idf. Sentences was scored accordingly, and the best sentences were returned. Python was used to run the analysis. Nltk, and sklearn packages were used.

Methodology

Corpus Creation

Data set was a collection of Business Insider articles for the years of 2013 and 2014. Specifically, each text file represented one article published in each day.

particular date.

Question Classification

Some labeled question data sets could be found on the internet. According to the features such Wh\_words(What, Who, When, Which, Whose, Where, etc), POS tag of the first and second words after those Wh\_words(Bigram, Trigram), the question was labeled as one of ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC and other features.

Document Retrieval

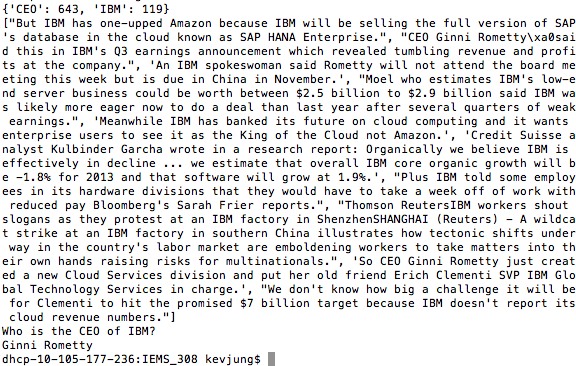
A subset of documents from corpus that contains any key words from the question was selected. The order of those documents was listed according to the value of TF-IDF value.

Answer with Rule Based Classification

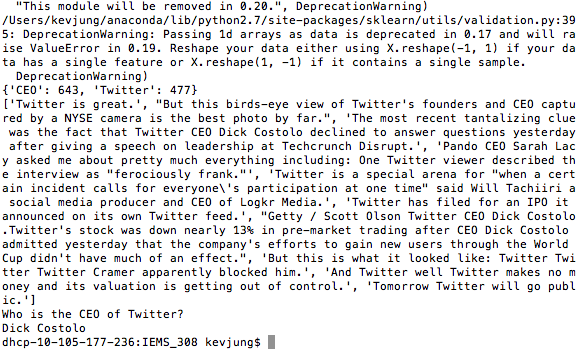
Out of those subsets of documents, sentences with related POS tagger to one of ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC labels were scored. It is arranged into descending orders of TF-IDF scores.

Analysis

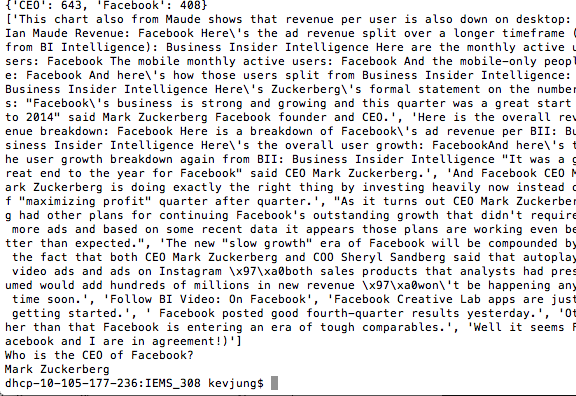
“Who is the CEO of IBM?”



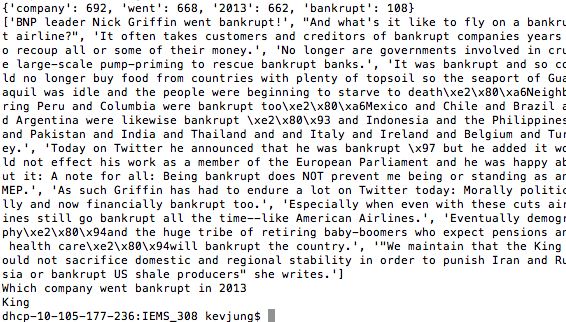
“Who is the CEO of Twitter?”



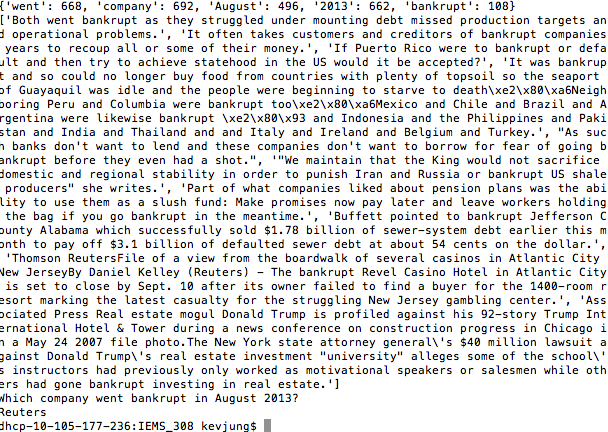
“Who is the CEO of Facebook?”



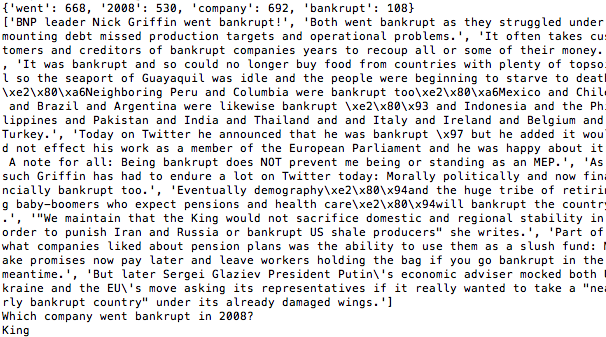
'Which company went bankrupt in 2013'



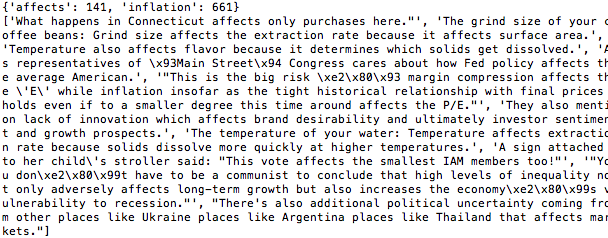
‘Which company went bankrupt in August 2013?’



Which Company Went bankrupt in 2008?



What affects inflation?



What affects GDP?

