Third Homework Assignment (Due March 2 at 10 am)

BusinessInsider is a portal for business news. The scraped articles for years 2013 and 2014 are available at https://www.dropbox.com/s/g43k1qzhpx3bwf5/2013.zip?dl=0 and https://www.dropbox.com/s/siuc4loq2fxsr5y/2014.zip?dl=0

There is one file per day. You are not allowed to use any other corpora for this homework assignment. You need to perform the following tasks.

1. Extract all company names from the files.
2. Extract all numbers involving percentages. Note that sometimes the corpus has "0.5%" and other times "point five percent." (and there might be other forms)
3. Extract all names of CEO's.

I anticipate that supervised learning will be used. For this you will have to create a training data set with labeled instances. Training labeled data is available in https://www.dropbox.com/s/yrkvnokw4ija2u9/labels.zip?dl=0

You are not allowed to use a third-party NER tool.

As deliverables you have to provide by March 2, 10 am on github:

1. Files with extracted entities (3 files)
2. A document specifying a list of the features used, the process taken (regex's used, other preprocessing techniques used, etc), the selected classification model, and the performance of your model.
3. The source code behind your NER classifier.

In the fourth homework assignment you will develop the answers for questions of the type

- Which company went bankrupt in January, 2013?
- What affects GDP? What percentage of drop or increase is associated with this property?
- Who is the CEO of company X?

(So when you do your entity extraction you might want to think about how it is going to help the last homework assignment.)