Aneesh Kudaravalli
IEMS 308
HW 3

Entity Extraction Process

Overview:

I started the extraction process by running the text data through the spaCy entity recognition system (https://spacy.io/usage/linguistic-features#named-entities). This uses a variety of statistical features to recognize entities, and I pulled all the name, organization, and percent entities from the output.

The percent outputs from spaCy are one of the assignment objectives, and spaCy locates all instances of “%”, “percent”, “percentage”, or “pct” in the data and accurately pulls the full statement regarding the amount of the percent (i.e. half a percent, more than 50 percent). The resulting percents found throughout both the 2013 and 2014 text data are aggregated in the finalfinalperclist.csv file.

To further classify names as CEOs, a logistic classification model was trained on the name entities found by spaCy in the 2013 text data and tested on the entities in the 2014 text data. This classification model looked at whether or not a name entity appeared in the given CEO labeled training data as the response indicating a name is a CEO. The binary features used to classify the names as CEOs are whether or not “CEO” or “chief executive” appears in the same sentence as the name entity and whether or not the name appears on a list of popular CEOs. The resulting CEOs classified from the model on both the test set and training set are aggregated in the finalfinalceolist.csv file. This model has a training misclassification rate of 2.0% and a test misclassification rate of 3.6%.

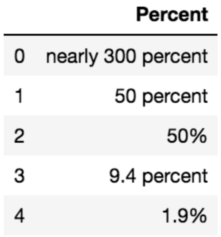
In a similar fashion, another logistic classification model was trained on the organization entities found in the 2013 data and tested on the 2014 data. Again, the response indicating whether an organization is a company is whether that organization name appears in the companies labeled training data. The features used to classify the orgs as companies are whether keywords (“stock”, “price”, “founder”, “partner”, “company”, “companies”) appear in the same sentence as the organization entity, the number of words in the organization entity, and whether the organization appears on the 2019 Forbes2000 list of companies. The resulting companies classified from the model on both the test set and training set are aggregated in the finalfinalcomplist.csv file. This model has a training misclassification rate of 8.5% and a test misclassification rate of 8.9%.

** All of the data pre-processing and feature building is done in the attached Python file. All of the logistic classification training and organizing the final deliverables is in the attached R file.

Processing and Feature Selection:

Percents:

As mentioned earlier, there was no need to conduct supervised learning on percents as spaCy accurately identifies all cases where "%" or "percent" appears in the dataset along with the surrounding context information around the %:



	Percent
0	nearly 300 percent
1	50 percent
2	50%
3	9.4 percent
4	1.9%

Since the classifier is locating all instances of percents in the text, no further classification or validation with the labeled set was needed. The resulting percents found throughout both the 2013 and 2014 text data are aggregated in the finalfinalperclist.csv file.

Names to CEOs:

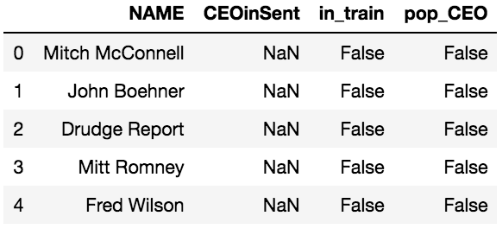
The main feature used to classify names as CEOs is whether "CEO" or "chief executive" (case insensitive) appears in the same sentence as the identified name. Other keywords such as "founder" or "partner" were considered but ultimately discarded as they would tend to include names that were not CEOs of companies, but rather founders of startups or partners of certain firms.

When testing this feature, I noticed that that many of the most famous CEOs – Mark Zuckerberg, Warren Buffett, etc. don't have "CEO" appear in sentence as its common knowledge that they are CEOs. Thus, I included another feature to indicate if entity appears on the Wikipedia list of CEOs as this is a small group of very popular CEOs and it covers the people famous enough to not need the CEO label:
https://en.wikipedia.org/wiki/List_of_chief_executive_officers

I also removed all names that are only one word long as they tend not to contain enough information to correctly identify a CEO. Furthermore, usually in news articles these one word occurrences of names appear after the full name was stated earlier.

I also removed most of the bad data where an input error makes it so that special characters are within the name.

The resulting dataset that the logistic classification model was applied to, looked like this where "in_train" is the response indicating whether an entity appears in the labeled training list:



	NAME	CEOinSent	in_train	pop_CEO
0	Mitch McConnell	NaN	False	False
1	John Boehner	NaN	False	False
2	Drudge Report	NaN	False	False
3	Mitt Romney	NaN	False	False
4	Fred Wilson	NaN	False	False

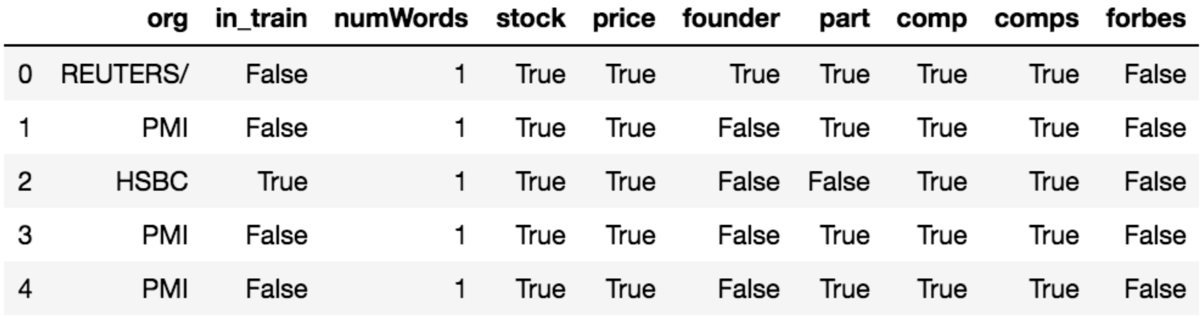
Orgs to Companies:

Since Business Insider tends to report information about founders of companies and about prices and financial info of companies, I included features that indicate whether or not these keywords (case insensitive) appear in the same sentence as the entity anywhere throughout the dataset: “stock”, “price”, “founder”, “partner”, “company”, “companies”.

Another thing I noticed during data exploration is that many of the organizations made up of multiple words are not companies. These are usually smaller organizations while companies tend to have shorter, recognizable names. So, I included number of words in the entity as another feature.

Running logistic regression at this point and removing duplicates resulted in a very small dataset of around 250 companies. Thus, I included a feature to indicate whether or not an organizations exists in the 2019 Forbes 2000 list of largest companies to classify the organization (https://www.forbes.com/global2000/). I did this for the same reason as including the pop_CEO feature – it increased the number of organizations predicted to be a company with an accurate feature, and it does not dominate the model (it does not discard all companies not on the Forbes 2000 list).

The resulting dataset that the logistic classification model was trained on looks like this where “in_train” is the response indicating whether an entity appears in the labeled training list:



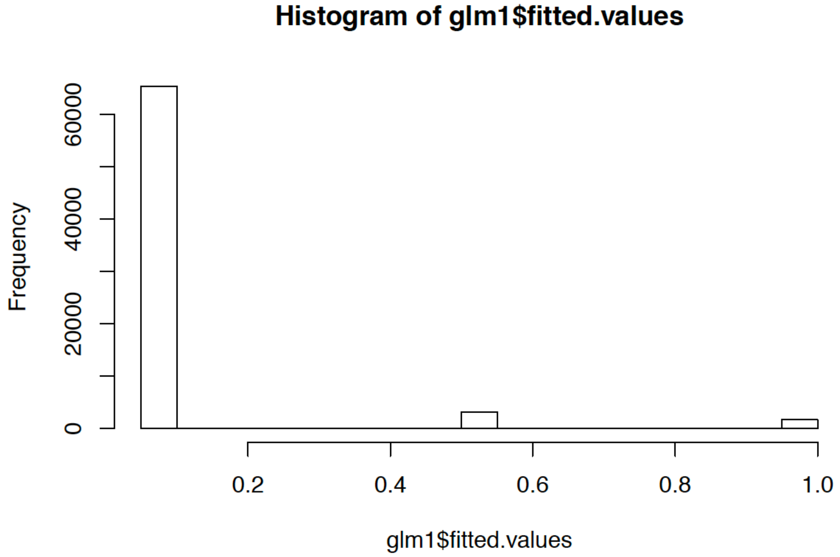
	org	in_train	numWords	stock	price	founder	part	comp	comps	forbes
0	REUTERS/	False	1	True	True	True	True	True	True	False
1	PMI	False	1	True	True	False	True	True	True	False
2	HSBC	True	1	True	True	False	False	True	True	False
3	PMI	False	1	True	True	False	True	True	True	False
4	PMI	False	1	True	True	False	True	True	True	False

Logistic Classification Model and Results:

Names to CEOs:

A logistic classification model was applied to the 2013 name entities that were cleaned and had features built out. The resulting coefficients on the model and histogram of fitted values are shown here:

```
##
## Call:  glm(formula = in_train ~ as.logical(CEOinSent) + as.logical(pop_CEO),
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##              (Intercept)  as.logical(CEOinSent)TRUE
##                   -2.328                      2.494
##   as.logical(pop_CEO)TRUE
##                    6.535
##
## Degrees of Freedom: 70187 Total (i.e. Null);  70185 Residual
## Null Deviance:       54380
## Residual Deviance: 43710     AIC: 43720
```



Using this information, I set the classification boundary to .4, so that if the probability of a name being a CEO is greater than .4, then the model considers this name to be a CEO. In the context of this model, this essentially means that if either the CEOinSent feature is true, or if the pop_CEO feature is true, then that entity is predicted to be a CEO.

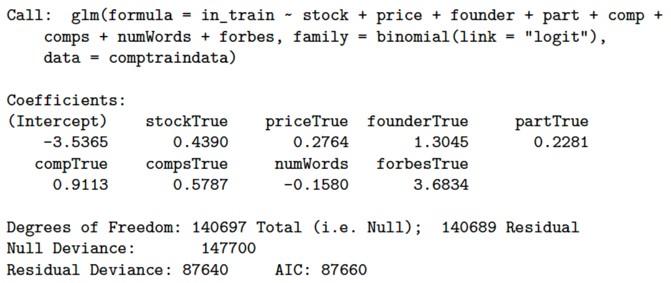
To calculate the training misclass rate, I looked at the percentage of time where the fitted value from the classification was greater than .4, but that name entity did not appear in the labeled training data. **The training misclass rate is 2.0%**

This model was then applied to the 2014 cleaned name entities. The fitted values showed a similar distribution to the previous 2013 data. Using the same boundary and formula from the training data, **the test misclass rate was calculated to be 3.6%**. Considering the low misclass rate of the model applied to test data, this seems to be a pretty accurate way of classifying CEOs among names in the Business Insider data.

The resulting CEOs classified from the model on both the training set and test set are aggregated in the finalfinalceolist.csv file with duplicates removed.

Orgs to Companies:

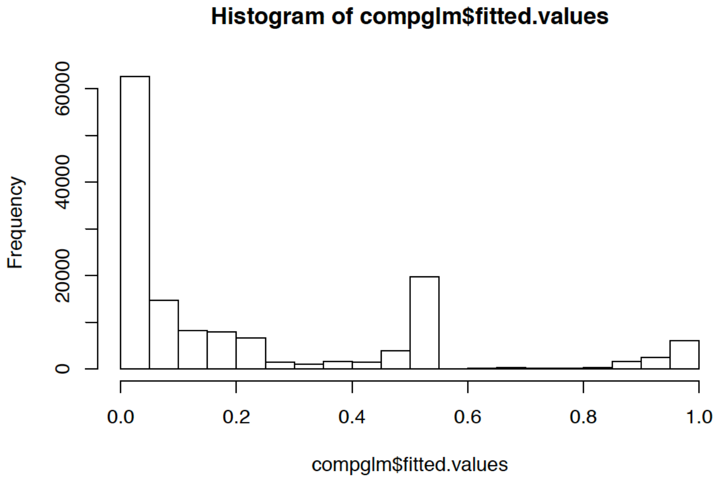
A logistic classification model was applied to the 2013 organization entities that were cleaned and had features built out. The resulting coefficients on the model and histogram of fitted values are shown here:



```
Call:  glm(formula = in_train ~ stock + price + founder + part + comp + 
    comps + numWords + forbes, family = binomial(link = "logit"), 
    data = comptraindata)

Coefficients:
(Intercept)    stockTrue    priceTrue  founderTrue     partTrue  
    -3.5365       0.4390       0.2764       1.3045       0.2281  
   compTrue    compsTrue     numWords   forbesTrue  
     0.9113       0.5787      -0.1580       3.6834  

Degrees of Freedom: 140697 Total (i.e. Null);  140689 Residual
Null Deviance:	    147700 
Residual Deviance: 87640 	AIC: 87660
```



By setting the classification boundary to 0.5, the predicted number of companies (30,813) is very close to the number of organizations that appear in the training labeled data (30,717). So, any organization that had a predicted probability of greater than .5 was classified as a company.

To calculate the training misclass rate, I looked at the percentage of time where the fitted value from the classification was greater than .5, but that organization entity did not appear in the companies labeled training data. **The training misclass rate is 8.5%**

This model was then applied to the 2014 cleaned organization entities. The fitted values showed a similar distribution to the previous 2013 data. Using the same boundary and formula from the training data, **the test misclass rate was calculated to be 8.9%**.

These misclass rates on organizations to companies are slightly bigger than the ones for names to CEOs. That said, removing duplicates from the final classified companies significantly decreases the size of the companies lists from around 60,000 to 440. This leads us to believe we have indeed found most of the companies included in the data, but more tuning could be done to expand the amount of organizations classified as companies.

The resulting companies classified from the model on both the training set and test set are aggregated in the finalfinalcomplist.csv file with duplicates removed.