Aneesh Kudaravalli

IEMS 308

HW 3

**Entity Extraction Process**

**Overview:**

* Used Spacy to identify entities initially
* Spacy identified Names, Organizations, and Percents
* Supervised learning was conducted on the identified Names and Organizations to further classify them as CEOs or Companies respectively
  + To determine if a name or organization is indeed a CEO or Company, I checked to see if that name or organization appears in the labeled training data we were given
  + 2013 articles were used as a training set (model is predicting correctly if it identifies a name as a CEO and that name is in labeled training data we were given) or if it identifies an org as a company and that org appears in the labeled training data we were given
  + The trained model was then used to predict on the 2014 data,
  + Misclass training and test rates were calculated based on appearance in labeled training data
* 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 %
* Final files include (duplicates removed)
  + – all percents captured by spacy on 2013 and 2014 data.
  + Names predicted to be CEOs from 2013 training set + Names predicted to be CEOs from 2014 test set
  + Orgs predicted to be companies from 2013 training set + orgs predicted to be companies from 2014 test set

**Processing and Feature Selection:**

**Names to CEOs:**

* + Main feature is whether “CEO” or “chief executive” (case insensitive) appears in the same sentence as the identified name
  + Noticed that many of the most famous CEOs – Elon Musk, Warren Buffett, etc. don’t have “CEO” appear in sentence as its common knowledge
    - Included another feature if entity appears on the Wikipedia list of CEOs (as this is a group of very popular CEOs)
  + Also removed all names that are only one word long (just last names or just first names) as it is not enough information and usually these one word occurrences of names appear after the full name was stated earlier
  + Removed any bad data (name is not spelled properly, or input error)

**Orgs to Companies:**

* + First feature is whether that company exists in the Forbes 2000 list of largest companies
  + Second feature is how many words are in the organizations – I noticed that many of the organizations that are not companies have multiple words
  + Removed any bad data (name is not spelled properly, or input error)

**Logistic Regression Model and Results:**

Training Set (2013) Misclass rate on identifying **orgs as companies** (misclassified if probability is greater than .5, but data does not appear in labeled training data): 0.007370396

Test Set (2014) Misclass rate on identifying **orgs as companies** (misclassified if probability is greater than .5, but data does not appear in labeled training data):

0.01459092

The code for classifying **names as ceos** isn’t running properly in the latest iteration, but initial misclass rates for both test and train data were under 5%.