Kathleen Young
IEMS 308, Professor Klabjan
7 March 2018

Named Entity Recognition for Business Insider Articles

List of features:

1. Features of the word:
    a. Last three letters
    b. Binary all uppercase
    c. Binary starts with a capitalized letter
    d. Binary is digit
    e. POS tag
    f. First two letters of POS tag (indicates the main part of speech, such as verb, noun, etc.)
    g. Binary beginning of sentence
    h. Binary end of sentence
2. Features of the previous and next word:
    a. Binary starts with a capitalized letter
    b. Binary all uppercase
    c. POS tag
    d. First two letters of POS tag

Process:

1. CEO and company CRF model:
    a) Used nltk package to sentence tokenize, word tokenize, POS tag, noun phrase chunk, and IOB tag
        i. Noun phrase chunk used for IOB: {<NNP>{1,}}. This tags one or more proper nouns in a row, which should capture almost all CEO names and company names. This is the jumping off point for tagging the corpus with the provided training data.
    b) I looped through the IOB tagged corpus and replaced the "B-NP" and "I-NP" tags with "CEO" or "COMPANY" if they existed in the training data. Everything else is labeled "O."
    c) Apply function to assign features and get labels
    d) Train the CRF model on the corpus
    e) Use the trained CRF model to make predictions on the corpus
    f) Create final extracted output lists by combining compound tags
        i. The model can only tag single words as companies or CEOs. If the model tagged two or more words in a row as CEO, it likely indicates a first and last name. The same goes for multi-word company names. If two or more tags have consecutive indices, they are combined to appear as one item on the final extracted output list, not two or more.
2. Percent regular expressions:
    a. Get the entire corpus as a single string
    b. Instead of training a model to recognize percentages, I used a rules-based method. The formats of percentages are highly predictable, so I wrote regular expressions to cover all representations of percentages in the training data:

i. [0-9-+]+?[0-9.]+%
ii. [0-9-+]+?[0-9.]+\spercent
iii. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?one\spercent
iv. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?two\spercent
v. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?three\spercent
vi. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?four(?:teen)?\spercent
vii. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?five\spercent
viii. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?six(?:teen)?\spercent
ix. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?seven(?:teen)?\spercent
x. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?eight\spercent
xi. (?:twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety)?-?nine(?:teen)?\spercent
xii. ten\spercent
xiii. eleven\spercent
xiv. twelve\spercent
xv. thirteen\spercent
xvi. fifteen\spercent
xvii. eighteen\spercent
xviii. twenty\spercent
xix. thirty\spercent
xx. forty\spercent
xxi. fifty\spercent
xxii. sixty\spercent
xxiii. seventy\spercent
xxiv. eighty\spercent
xxv. ninety\spercent
xxvi. one-?\s?hundred\spercent
xxvii. nineteen\spercent

c. Output extracted percentages as a list

Classification model:

- CRF, or Conditional Random Fields, is a model commonly used for named entity recognition. CRF models do well in prediction tasks where sequence is important. The model takes contextual information and state of the neighbors to predict the current condition, making it a good fit for predicting word tags based on the features of the word itself as well as the features of surrounding words.

Model evaluation:

- According to the report below, when the model predicts a CEO or company name, it is often correct—80% of the time for CEOs and 81% of the time for company names. However, the recall

statistics show that there are many named entities that the model overlooks. This is reflected in the f1-score as well.

	precision	recall	f1-score	support
CEO	0.80	0.01	0.01	195782
COMPANY	0.81	0.02	0.05	138828
O	0.98	1.00	0.99	16492032
micro avg	0.98	0.98	0.98	16826642
macro avg	0.86	0.34	0.35	16826642
weighted avg	0.98	0.98	0.97	16826642

- Here are the top predicting features for each label:

y=CEO top features		y=COMPANY top features		y=O top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+0.768	postag:NNP	+1.547	postag:NNP	+2.304	bias
+0.766	word.istitle()	+0.383	postag[:2]:NN	+1.902	postag:NN
+0.351	word[-3:]:erg	+0.336	word.lower():apple	+1.717	postag[:2]:VB
+0.344	word[-3:]:son	+0.307	word.istitle()	+1.686	BOS
+0.325	word.lower():bloomberg	+0.294	word[-3:]:gan	+1.499	EOS
+0.259	+1:postag:VBD	+0.293	word[-3:]:ple	+1.382	postag:.
… 26394 more positive …		+0.261	word.lower():reuters	+1.382	postag[:2]:.
… 6893 more negative …		*… 7844 more positive …*		+1.360	postag[:2]:IN
-0.257	+1:postag:NN	*… 5043 more negative …*		+1.360	postag:IN
-0.533	word.isupper()	-0.576	-1:postag[:2]:NN	*… 199042 more positive …*	
-0.678	BOS	-0.801	BOS	*… 17639 more negative …*	
-1.201	bias	-1.103	bias	-2.550	postag:NNP

  - For CEO names, being a proper noun (POS tag NNP) or a title (first letter capitalized) are huge predictors, which makes sense for the names of a person. Being at the beginning of a sentence is strongly negatively correlated with being a CEO name.
  - For company names, being a proper noun or a noun at all are strong predictors for being a company name. Apple is clearly mentioned a lot in these Business Insider articles, because the word being “Apple” is also strongly correlated with being a company name. Being at the beginning of a sentence, or the previous word being a noun, are negatively correlated with being a company name.