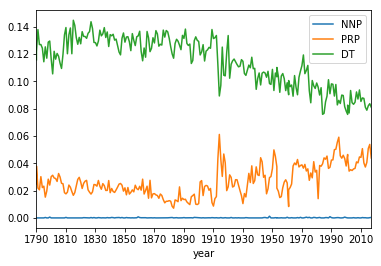
Assignment 3 - POS Tagging and Named Entity Recognition

# POS Tagging

1. Did in the code
2. Plot of the normalized NNP, PRP and DT for each of the speeches:  
   
3. The most common bigrams in the corpus are:  
   (('IN', 'DT'), 118368), (('DT', 'NN'), 113976), (('NN', 'IN'), 108712), (('JJ', 'NN'), 74781),   
   (('DT', 'JJ'), 62398)  
   The most common trigrams in the corpus are:  
   (('IN', 'DT', 'NN'), 59411), (('DT', 'NN', 'IN'), 52939), (('NN', 'IN', 'DT'), 47051),   
   (('DT', 'JJ', 'NN'), 38477), (('IN', 'DT', 'JJ'), 34132)
4. The top pos n-grams of some of the presidents:  
   pos ngrams for president George Washington

unigrams: [(('NN',), 2758), (('IN',), 2710), (('DT',), 2248), (('JJ',), 1420), (('NNS',), 1251)]

bigrams: [(('IN', 'DT'), 1270), (('NN', 'IN'), 1108), (('DT', 'NN'), 1099), (('DT', 'JJ'), 612), (('JJ', 'NN'), 591)]

trigrams: [(('DT', 'NN', 'IN'), 608), (('IN', 'DT', 'NN'), 608), (('NN', 'IN', 'DT'), 527), (('IN', 'DT', 'JJ'), 390), (('DT', 'JJ', 'NN'), 363)]

pos ngrams for president Abraham Lincoln

unigrams: [(('NN',), 4651), (('IN',), 4332), (('DT',), 3603), (('JJ',), 2616), (('NNS',), 1881)]

bigrams: [(('IN', 'DT'), 1906), (('DT', 'NN'), 1741), (('NN', 'IN'), 1656), (('JJ', 'NN'), 1139), (('DT', 'JJ'), 1029)]

trigrams: [(('IN', 'DT', 'NN'), 938), (('DT', 'NN', 'IN'), 828), (('NN', 'IN', 'DT'), 743), (('DT', 'JJ', 'NN'), 601), (('IN', 'DT', 'JJ'), 557)]

pos ngrams for president Richard Nixon

unigrams: [(('NN',), 3581), (('IN',), 2794), (('DT',), 2448), (('JJ',), 1646), (('NNS',), 1315)]

bigrams: [(('DT', 'NN'), 1172), (('IN', 'DT'), 1106), (('NN', 'IN'), 1054), (('JJ', 'NN'), 781), (('DT', 'JJ'), 637)]

trigrams: [(('IN', 'DT', 'NN'), 521), (('DT', 'NN', 'IN'), 445), (('NN', 'IN', 'DT'), 414), (('DT', 'JJ', 'NN'), 396), (('IN', 'DT', 'JJ'), 307)]

pos ngrams for president Ronald Reagan

unigrams: [(('NN',), 9464), (('IN',), 5893), (('DT',), 5078), (('JJ',), 4348), (('NNS',), 3532)]

bigrams: [(('DT', 'NN'), 2425), (('NN', 'IN'), 2156), (('JJ', 'NN'), 2035), (('IN', 'DT'), 1881), (('DT', 'JJ'), 1373)]

trigrams: [(('IN', 'DT', 'NN'), 920), (('DT', 'JJ', 'NN'), 886), (('DT', 'NN', 'IN'), 818), (('NN', 'IN', 'DT'), 686), (('NN', 'IN', 'NN'), 580)]

pos ngrams for president Barack Obama

unigrams: [(('NN',), 9171), (('IN',), 5970), (('DT',), 5061), (('NNS',), 4256), (('JJ',), 3998)]

bigrams: [(('DT', 'NN'), 2537), (('NN', 'IN'), 1752), (('IN', 'DT'), 1747), (('JJ', 'NN'), 1706), (('TO', 'VB'), 1336)]

trigrams: [(('IN', 'DT', 'NN'), 877), (('DT', 'JJ', 'NN'), 774), (('DT', 'NN', 'IN'), 695), (('NN', 'IN', 'DT'), 471), (('IN', 'DT', 'JJ'), 460)]

pos ngrams for president Donald J. Trum

unigrams: [(('NN',), 892), (('IN',), 563), (('DT',), 462), (('JJ',), 457), (('NNS',), 407)]

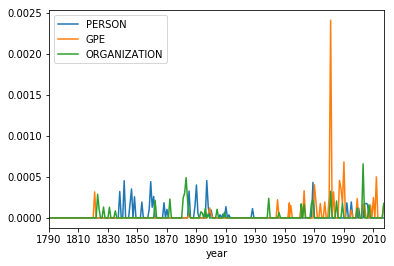
bigrams: [(('DT', 'NN'), 201), (('JJ', 'NN'), 199), (('NN', 'IN'), 180), (('IN', 'DT'), 163), (('NN', '.'), 149)]

trigrams: [(('DT', 'JJ', 'NN'), 79), (('DT', 'NN', 'IN'), 67), (('IN', 'DT', 'NN'), 66), (('IN', 'DT', 'JJ'), 58), (('NN', 'IN', 'DT'), 49)]

1. We can see a small difference in the unigrams, a switch between the JJ-adjective and NNS-noun plural in the two presidents.  
   We can also see that for the bigrams, minor switches. but i do notice that trump uses relatively a lot of adjectives and nouns compared to the other presidents.  
   In the trigrams there are more differences between the presidents, obviously, as there are more combinations. we can see that over the years the trigram of DT,JJ,NN is getting more popular.  
   My conclusion is that over the years the speeches content try to show more active approaches and unity in the nation.
2. The sampled sentences and tags are in the code.

I manually tagged the sentences and if that’s the ground truth than i could not find any major difference in the quality of the POS tagging. Their are both pretty good, still have some errors switching between Nouns and Verbs but overall it is good.

# Named Entity Recognition

1. Plot of the normalized GPE, PERSON and ORG NERs for each of the speeches:  
   
2. The changes I can see are that as the USA became a world leader there are more notifications of GPEs as the president of a leading country should address the rest of the world as well.
3. The most frequent GPEs:   
   ('u.s.', 99),

('d.c.', 6),

('u.s.-soviet', 5),

('austro-hungarian', 5),

('indo-china', 5),

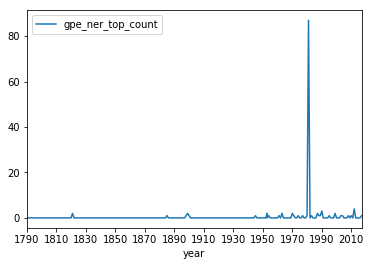
('iran-iraq', 2),

('u.s.a.', 2),

('st.', 2),

('american-made', 1),

('russia', 1)

1.   
   We can see the most of the references of GPEs were in the eighties, specifically on 1981 when there a great recession therefore there was a great concern about the state on the nation. So, a lot of u.s. mentioned in the text.
2. The sampled sentences and tags are in the code. There were not tags for any of Washington’s speeches. There were few for Clinton, but I did not extract them specifically as there won’t be sampling in that case.

I manually tagged the sentences and if that’s the ground truth than, first, i have to say that the NER tagger is **BAD** and doing awful job. As i said, the tagger still found some NER tags for Clinton i can say there’s a difference. But, it did not find many other tags for both for example:

In Washington’s:

* Senate
* United States – divided
* Virginia
* Congress

In Clinton’s:

* America
* Federal courts – divided

My conclusion is that the tokenization of the words in the sentence is bad. As I believe there’s no chance that some of these tags would not been recognized if they were in the same token.