Sunday, November 16, 2014

NLP Milestone

1.

We made some simple changes to start code and we have our F score to be 0.16

Precision is 62/129, which means we have made 129 name predictions and 62 of them are correct, Recall is only 62/647, witch means there are totally 647 names in gold standard and we only found 62 of them and made the prediction, this number is the main reason that we have such small F score and we then will focus on how to improve this.

2. After taking a look at tag predictions and gold standard, I noticed that we get good predictions for the words or phrase that are totally the same. This is reasonable considering that our codes right now only take the words around one certain word to be its attributes, and we assign them weights regarding the distance. But by doing this way it is hard to predict for unseen word in a new file. So we need to add new attributes to analyze structure of the sentences.

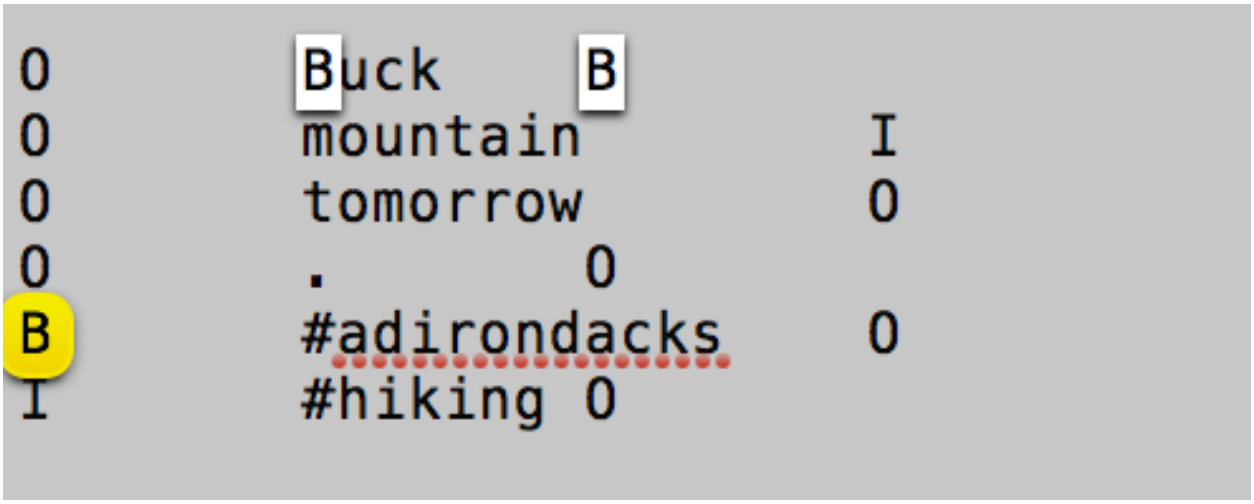
3. In model, we have FILEHEADER, LABELS, ATTRIBUTES, TRANSITIONS, STATE_FEATURES

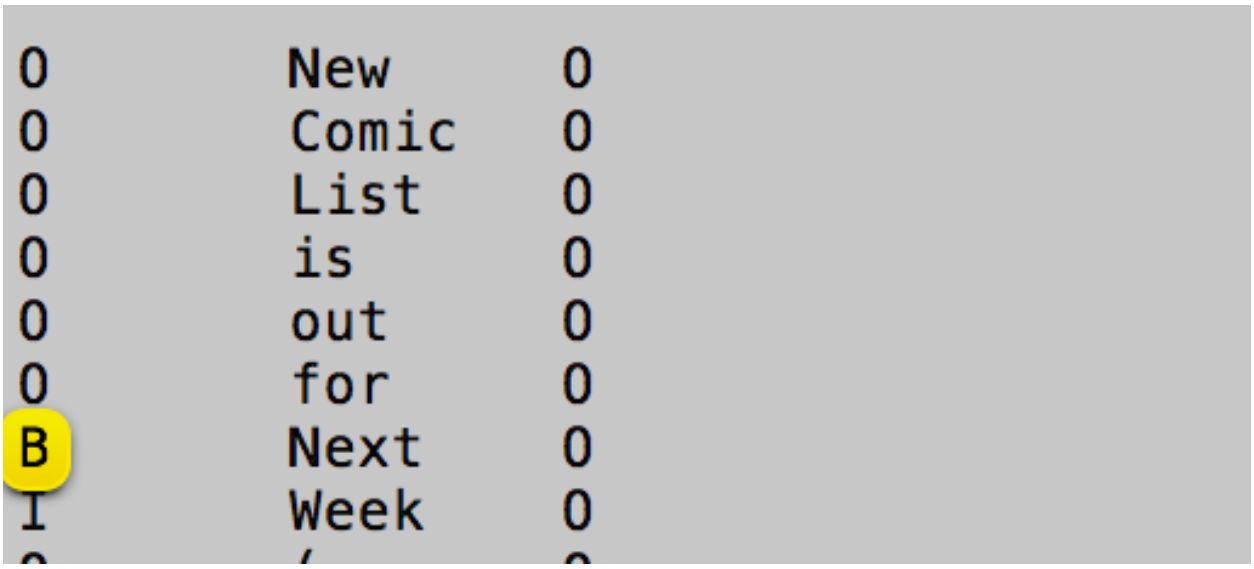
example of high weights:

word[0]=M --> B: 0.687369
word[0]=the --> O: 1.024825
__BOS__ --> I: 0.614590
word[1]='s --> B: 0.746043

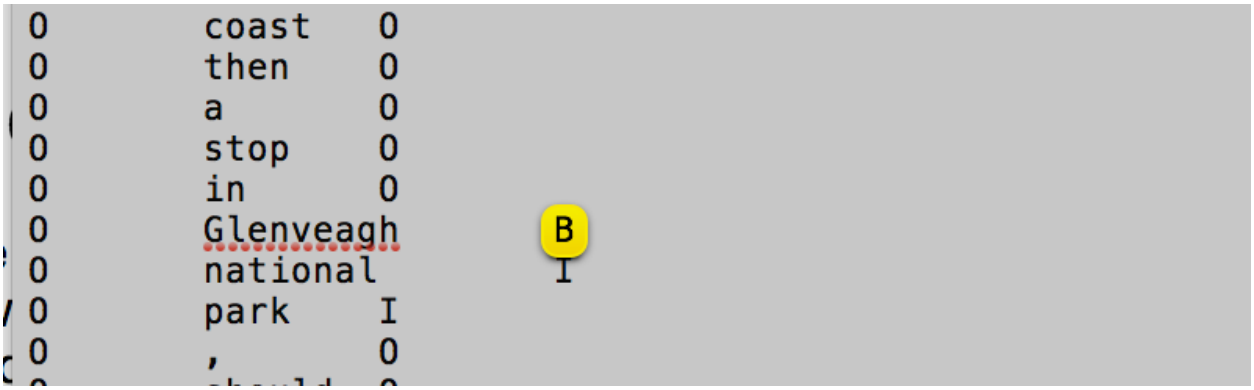
the problem is just I mentioned above, we only take word and relative position into consideration and this makes prediction bad

4.
Two example of False Positive:





Example of False Negative:



For this I was considering to add an attributes to check if the word starts with a capital letter