Homework 4- Named Entity Recognition Report

Group: Avenger

Group members: Zhenqi Li, Zhiren Chen, Zhengwu Yuan

1. We used the Sequence learning toolkit, from https://github.com/larsmans/seqlearn.
2. This is a sequence classification toolkit for Python, and it is designed to extend scikit-learn and offer as similar as possible an API.
3. I first do some preprocess of my input training set and testing set:

   For the training set: I delete the line number because load_conll() function from the toolkit just splits off the tag at the end of the line for each training example. And it returns the rest of the line as a string. If we don't delete the line number, then it will make each word "feature" distinct, and we can not learn anything from a feature like that.

   For the testing set: for the same reason as the training set, we delete the line number, and also we add an dummy last column tags, because the the load_conll() function assumes the last column is where the tags are, so I add the last column with all "O"s.
4. For the feature(sentence, i) function, I use different shape of words as the features, such as uppercase and lowercase and the length of string. And I also add the features of the previous three and the three words after the entity to the current words to make the prediction more accurate.
5. Finally, in the main function, I get the model from the StructuredPercptron() function and I set the iteration times as 30 to increase the F1-measure score. Then I get the y_pred, which is the final tags for entities. Then format the out-put to the .txt file.