HW3

SHENGJIE LIU

1. For CEO problem:

In this problem, feature used: the length of first name, the length of last name, the first name index, the last name index, the before\_first\_name\_word pos\_tag, the after\_last\_name\_word\_tag, the first name tag, the last name tag, the before\_first\_name\_word Capitalization, the after last\_name\_word Capitalization.

Based on this feature selection, we can get our confusion matrix towards this model.

Confusion Matrix:

[[46697 5786]

[ 1659 989]]

Accuracy: 0.8649580091055848

Precision: 0.14597785977859778

Recall: 0.3734894259818731

F1 Score: 0.20991191764830733

1. For companies problem :

In this problem, feature used: the length of the first word, the length of the last word, the first word tag, the last word tag, the percentage of Capitalization in a company name, whether it can have some keyword such as Corp, Inc, and so on.

Based on this feature selection, we can get our confusion matrix:

Confusion Matrix:

[[38940 4323]

[ 2203 1100]]

Accuracy: 0.859932003023342334

Precision: 0.9012313213213231

Recall: 0.1231323131321321

F1 Score: 0.92274271332131

1. For percentage problem, due to time reason , I just come up with the re in this model. [0-9] %|[a-z]+ percent|[A-Z][a-z]\* percent|[0-9] percent