HW1 - Asaf Mizrahi & Yahel Sherman

In this first assignment we were tasked to solve the NER task using manual features.

In order to solve this task we tried a couple of features with the data we had available. The features we tried and used solving this task were:

Capitalization – Checks if the first letter of the word is capitalized (isupper)

Vocabulary – Using the words in "train.txt" and whether they are entities or not.

Speech Tag – Using the words speech tags (JJ,NNP,VB, etc..) in train.txt and whether they are entities or not.

We also tried a couple of other features that decreased our score on the train and eval texts, so in the final version we decided to not use them:

First Word – Checks if the word is the first one in its sentence.

Days – We concluded from train.txt that the days of the week can't be entities.

Months – We concluded from train.txt that the months of the year can't be entities.

ing – Words that end in 'ing' are progressive verb tenses or nouns, either way they can't be entities.

We picked the features for the competitive part by trial and error. We trained the model using different combinations of all the features we listed above and we got the best score on the train and eval texts by using only the first 3 features. In order to increase our training database we combined the train and eval txt provided into one training text.

The NER task is binary in its nature, is the word in the sentence is an entity or not. For this reason we used SVM model to solve the problem. In order to receive the best results we used the implemented "grid search" available in the model.

Final results of our model:

**Training accuracy: 0.991**

**Validation accuracy: 0.986**

Confusion matrix:

