**CE807-7-SP: Text Analytics**

**Assignment 1**

**Report**

**Name: Yaowei Lyu**

**Tutor: Dr habil Ansgar Scherp**

**Student Number: 1802697**

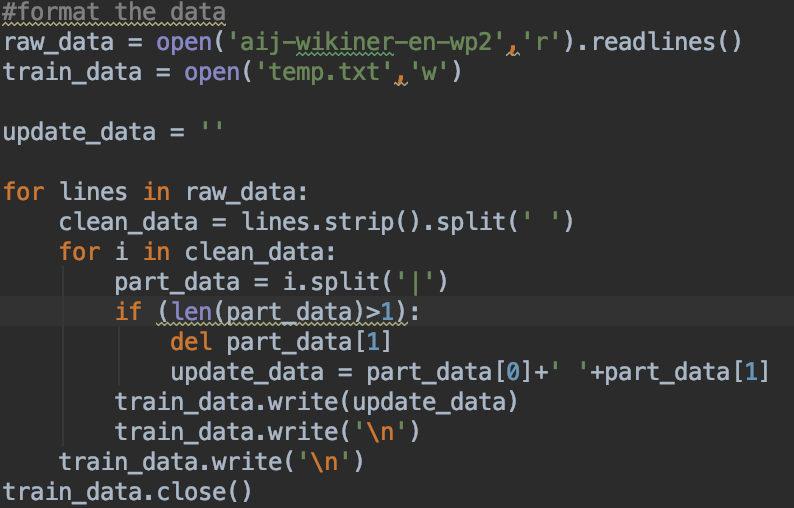
1. **Background**

In this assignment, we need to build a name entity recognition system and get the accuracy of the system. I suppose to use the perceptron model to complete this task, perceptron is a linear model that accept two vector as the input and get the a instance example as the output. There is a package in the SKLearn called perceptron that fit our requirements. We need to clean the data first and then fit the model. After doing that we, we can find the accuracy of the model.

1. **Reason for choosing Perceptron**

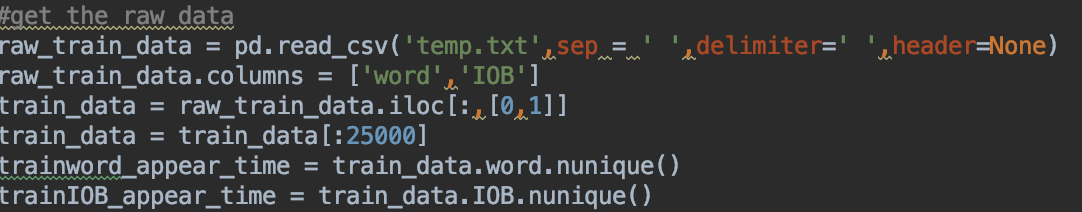
First of all, the perceptron is a linear model of dichotomies, which input is the eigenvector of the instance and the output is the category of the instance. It’s a discriminant model. The raw data we get from the Wikipedia after cleaning have two features: words and IOB format. So it fit for this model. Secondly, perceptron is going to find the parameter that get the minimum loss of the model. It can improve the performance of the model. For example, it get two parameter first then build a function to check whether fit for the data.

1. **Process** 
   1. **Format the data**

****

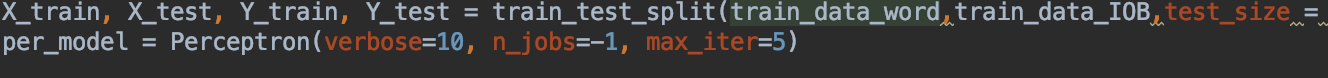
We can get the raw data from the aij-wikiner-en-wp2 by using read function. Then we create a file called temp to store the format data which remove the pos-tag.

* 1. **Get the raw data**



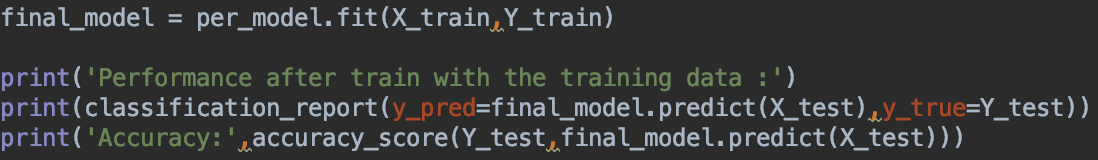
After formatting the data, we can get the data by read\_csv from the panda package. There is also another function in panda called iloc, it is going to select the data based on the index.

* 1. **Spilt the Data**

****

We can use the function train\_test\_split of sklearn package to split the data and get the train data set and the testing data set. Meanwhile we create a perceptron model by using the train data set.

* 1. Get the Results



After train the model fit with the train data set. We can get the accuracy of the model by using the accuracy\_score function of sklearn.

Results:

