NLP Task by
Mohamed Adel gomma Mohamed Elghannam

Documentation

1 - The data consists of two columns (job name , industry name)

2- the data is categorical so we will use encoding

3- the industry is the target column which consists of 4 labels IT , Marketing , Accountancy , Educational

4- because the data is categorical so we encoded the job names by using OneHotEncoder and we gave an int value for each value of the target column
"IT":0 , "Marketing":1 , "Education":2 , "Accountancy":3

5- The data is (imbalanced) biased towards IT in a very badway so we will remove duplicates and then use SMOTE to oversample the minority labels to be equal to each other at the end.

Now the data is ready for the model

6-

I used support Vector Classifier because it is very effective in multi class classification and I can also control the data points that can enter the margin (C argument) .

- SVM is more effective in high dimensional spaces(especially when using one hot encoding which have increased my dimensional spaces)

-SVM is effective in cases where the number of dimensions is greater than the number of samples.

7 - Industry_name function uses the trained model and the job name to produce the name of the predicted industry

8- finally I have made RESTful API service to deploy the function of the model (industry_name) in HTTP page http://127.0.0.1:5000/

where we will insert the job name and press on submit button to get the industry name.