**Job title Classification by industry**

(iNetworks)

**About:**

**This document is meant to answer the asked questions related to job title classification task.**

1. **Which techniques you have used while cleaning the data if you have**

**cleaned it?**

* I have loaded the data using pandas data-frames
* I checked the data shape and take a quick look into it
* I check the duplicates in the dataset and there was a lot of duplicates which I removed
* I check the job title column and removed the special characters, number, stop words and there were any errors from scrapping.
* After checking the classes, we found the imbalance of data in one class and i solved this imbalance using under-sample method for the majority class which was (IT)
* I did split the data into train and test data to start the training process
* I used two models (SVM, MNB) for the classification
* SVM model had better accuracy with the default params

1. **Why have you chosen this classifier?**

* I used tow classifiers knows as they are good and recommended to text classification problems
* NB is recommended algorithm as it depend on the bayes theory which is good for text classification
* SVM is an algorithm that determines the best decision boundary between vectors that belong to a given group (or category) and vectors that do not belong to it.

1. **How do you deal with (imbalance learning)?**

* There are many techniques you could use to overcome the imbalance issue, as we could use resampling methods (upsample, undersample) based on the case
* We could also manipulate the start weights of each class to avoid low accuracy

1. **How can you extend the model to have better performance?**

* Performance has been increased after doing two important operations which was removing the duplicates from the model and factorize the words to help the model analyze it.
* The performance could have been increased by trying other imbalance techniques to overcome the imbalance issue completely
* We could increase performance using ensample methods and model stacking to have better accuracy

1. **How do you evaluate your model?**

* I have used recall,percsion and f-1 score to evaluate the model, as accuracy is not the only indicator of the model is doing

1. **What are the limitations of your methodology or where does your approach fail?**

* Imbalance data was not handled in the best way, maybe with more data in the imbalance class it would give higher accuracy.