

Name: Mariam Mohamed Mahmoud.

Faculty: Graduated From Faculty of Computer and
Information Science Ain Shams University.

NLP Task Report

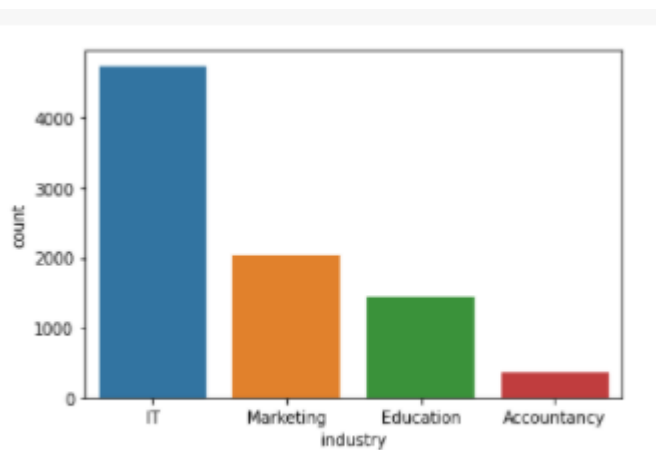
- Problem definition:

-given an **imbalanced** dataset regarding job industry, the main target is to predict the industry depends on the job title.

- Dataset columns:

	job title	industry
0	technical support and helpdesk supervisor - county buildings, ayr soa04086	IT
1	senior technical support engineer	IT
2	head of it services	IT
3	js front end engineer	IT
4	network and telephony controller	IT

- Dataset imbalanced:



As we can see, there is class imbalance here

- Data Pre-processing:

- 1- Loading the words needed to be removed using (stopwords.words).
- 2- Remove 'it' word from (stopwords.words) as it simulates (IT industry).
- 3- Using Regular expressions to replace special characters with space.
- 4- Lemmatizing the words in the dataset.
- 5-Pre-process and Vectorize train and test data

- Modelling:

1-applying Machine learning models on an **Imbalanced** Data:

- 1.1 : LogisticRegression accuracy = 0.935724266418258
- 1.2 : MultinomialNB accuracy = 0.9273404750815091
- 1.3 : SVC accuracy = 0.9385188635305077
- 1.4 : LinearSVC accuracy = 0.9417792268281323
- 1.5: RandomForest accuracy = 0.9315323707498836

2-applying Machine Learning models on **balanced** Data with technique

'RandomOverSampler': for making data balanced:

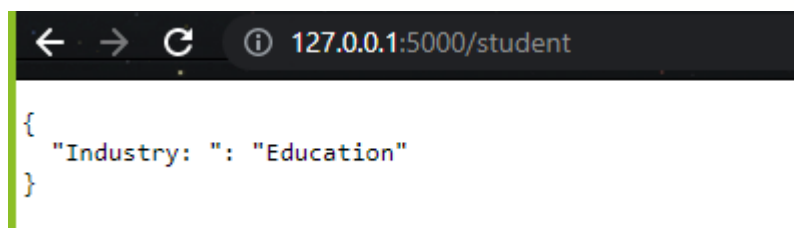
- 2.1: LogisticRegression accuracy = 0.9427107591988821
- 2.2: MultinomialNB accuracy = 0.9259431765253843
- 2.4: LinearSVC accuracy = 0.9385188635305077

3- applying Machine Learning models on **balanced** Data with technique

'SMOTE': for making data balanced:

- 2.1: LogisticRegression accuracy = 0.936655798789008
- 2.2: MultinomialNB accuracy= 0.9189566837447601
- 2.4: LinearSVC accuracy = 0.935258500232883

- Using Flask API:



```
{
  "Industry": "Education"
}
```

-Questions:

1- Answer Q1:

- 1-set all the words to lower case.
- 2-remove dash,() anything except spaces.
- 3-remove '£' and replace it with space
- 4-Remove any slash or digit.
- 5-remove the stopwords.
- 6-using lemmatizig concepts.

2-Answer Q2:

- 1-I used logistic regression model as it classifies the data better than other model , plus it fits this kind of data.

3-Answer Q3:

- 1-by using RandomOverSampler, SMOTE which is Randomly duplicate examples in the minority class.

4-Answer Q4:

- 1-the model could have better performance if the data had more features , to give the model the chance to learn form a big variety of info to be more generalized.

5-Answer Q5:

- 1-it depends on the problem in our case ,f1 score or accuracy will fit it as it's not a risky data

6-Answer Q6:

- 1-there's no enough data to make the model more generable

