Machine Learning Intern.

Job title Classification by industry

(Multi-text Text Classification Task)

Description:

You can think of the job industry as the category or general field in which you work. On a job application, "industry" refers to a broad category under which a number of job titles can fall. For example, sales is an industry; job titles under this category can include sales associate, sales manager, manufacturing sales rep, pharmaceutical sales and so on.

Details:

1- You are given a dataset that has two variables (Job title & Industry) in a csv format of more than 8,500 samples. (download link)
This dataset is imbalanced (Imbalance means that the number of data points available for different classes is different) as follows:

IT	4746
Marketing	2031
Education	1435
Accountancy	374

2- You are required to build a model using any Machine Learning classifier algorithm to classify job titles by the industry and provide us with insights on how your model works.

3- Answer the following questions:

- Which techniques you have used while cleaning the data if you have cleaned it?
- Why have you chosen this classifier? (E.g. I used Multinomial Naive Bayes because it is easy to interpret with text data and there are more than two outcomes).
- How do you deal with (Imbalance learning)?
- How can you extend the model to have better performance?
- How do you evaluate your model? (i.e. accuracy, F1 score, Recall)
- What are the limitations of your methodology or Where does your approach fail? (e.g. your predictions are biased because you do not have enough data for a certain class)

4- Create this script as a RESTful API service where the input is a HTTP request with a parameter for the "Job title" and the output is the expected industry.

Deliverables:

- Your source code.
- The report that explains your solution and answers the above questions.

Suggestions:

1. We prefer to use Jupyter Notebook with Python for easy documentation).
2. Use Flask API to deploy the RESTful API service.

Rules:

1. Describe your approach in detail in either Jupyter Notebook or word document including all your references and any assumptions you made.
2. Please include all the code snippets/ supporting files you have and organize in this clear folder structure
   a. Root Structure: YourName_IndustryClassificationTask
   b. Subfolders: 01_Code, 02_Documents