Project Idea :

The project idea is to recommend web articles for learners during their learning journey. Those articles will be recommended for the different nano degrees. e.g. Machine learning, Product management, UI/UX Design ... etc

Data set:

A JSON file containing 3 Categories [Engineering , Startups & Business, Product & Design]

And the goal is to classify the articles into these Categories

Steps:

1. Data exploration
2. Data Cleaning
3. Text Preprocessing
4. Model Training

data exploration:

- Investigating the target and the feature for additional information and deciding how to clean and preprocess the dat
- Count the target values
- Search for duplicates & empty fileds
- Create Word Cloud for each category

Data Cleaning :

- Drop duplicates.
- remove the empty fields.

Data Preprocessing :

- Remove punctuation
- Convert all texts to be in lowercase.
- Use nltk.tokenize for sentences tokenization.
- Remove stopwords from the tokenized text.
- Apply Lemmatization to the texts.
- DownSampling
- Save the final processed dataframe to be used in the next step of Model Training.

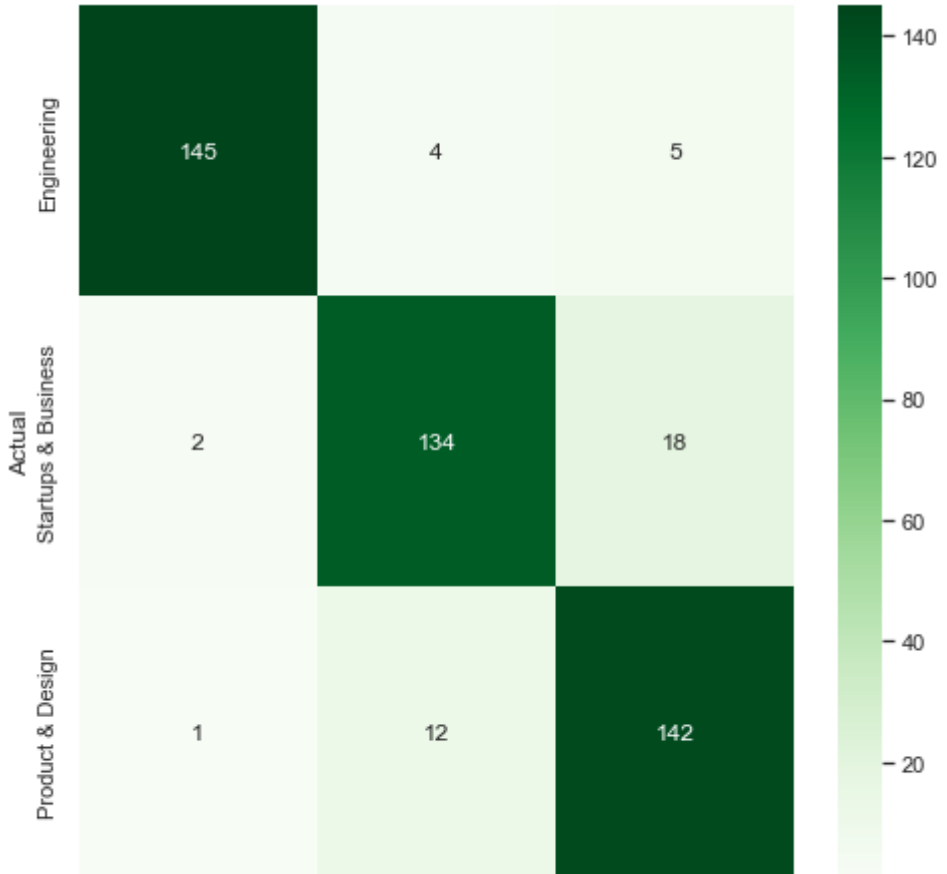
Model Training :

- MultinomialNB
- LogisticRegression
- LinearSVC

The data was good enough to be fitted to the used models and the results were greatly satisfying with the classical ML

91% accuracy

confusion_matrix



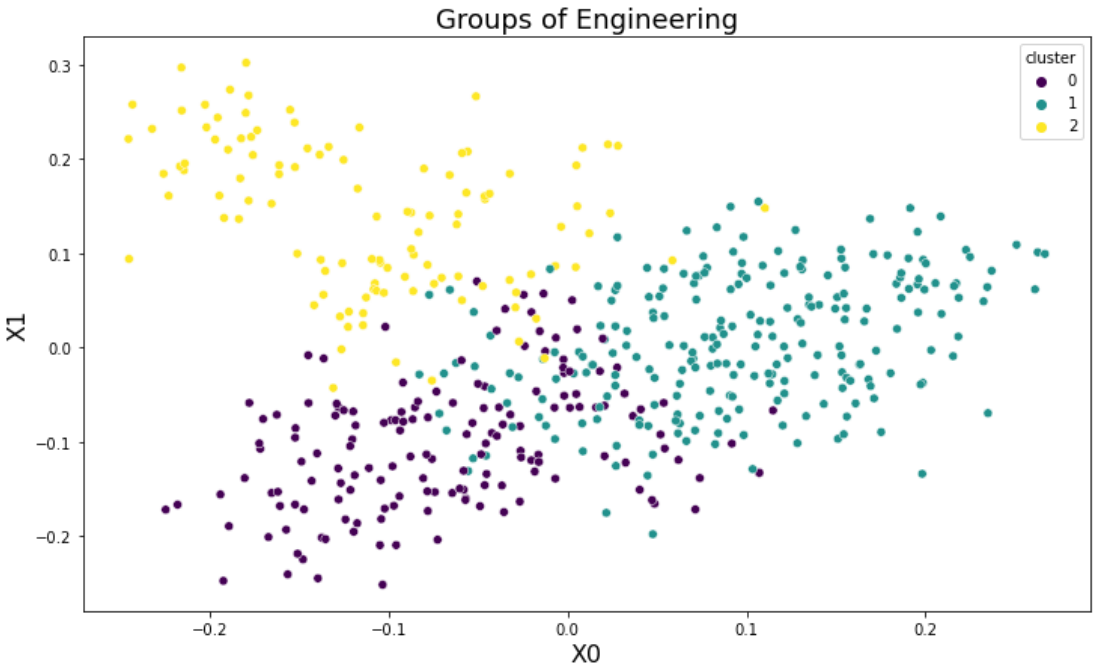
Further Improvement:

- Scarp more data to improve the model and create the Cluster
- Use a Deep Learning model like (LSTM) to increase the accuracy but first, we need a lot of data

Steps for Clustering each group :

1. I first subset each category
2. create Vectorize the body feature using TfidfVectorizer
3. cluster the category to subgroups using Kmeans
4. use PCA for Dimensional Reduction and Visualization
5. get the most relevant keywords for each group
6. Save the result in a JSON file for each category

Engineering Category :



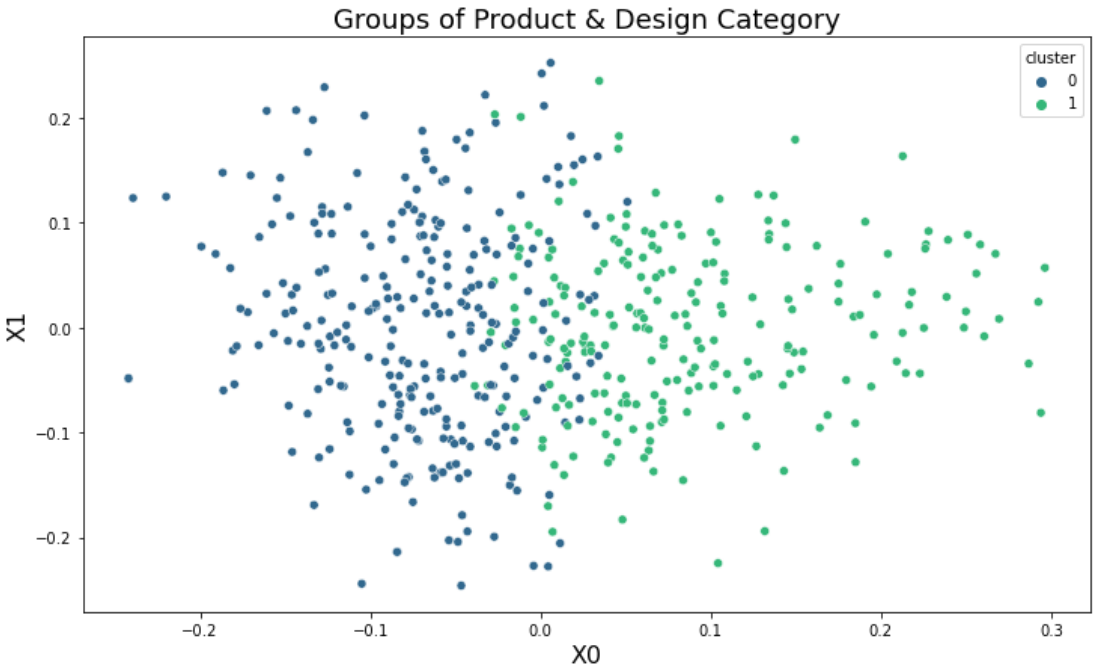

```
the keywords for each centroid of the KMeans

Group 0
time,need,use,function,object,like,app,react,javascript,code

Group 1
use,microservices,new,application,time,one,database,system,data,service

Group 2
trained,network,machine,data,deep,algorithm,neural,model,training,learning
```

Product & Design Category :

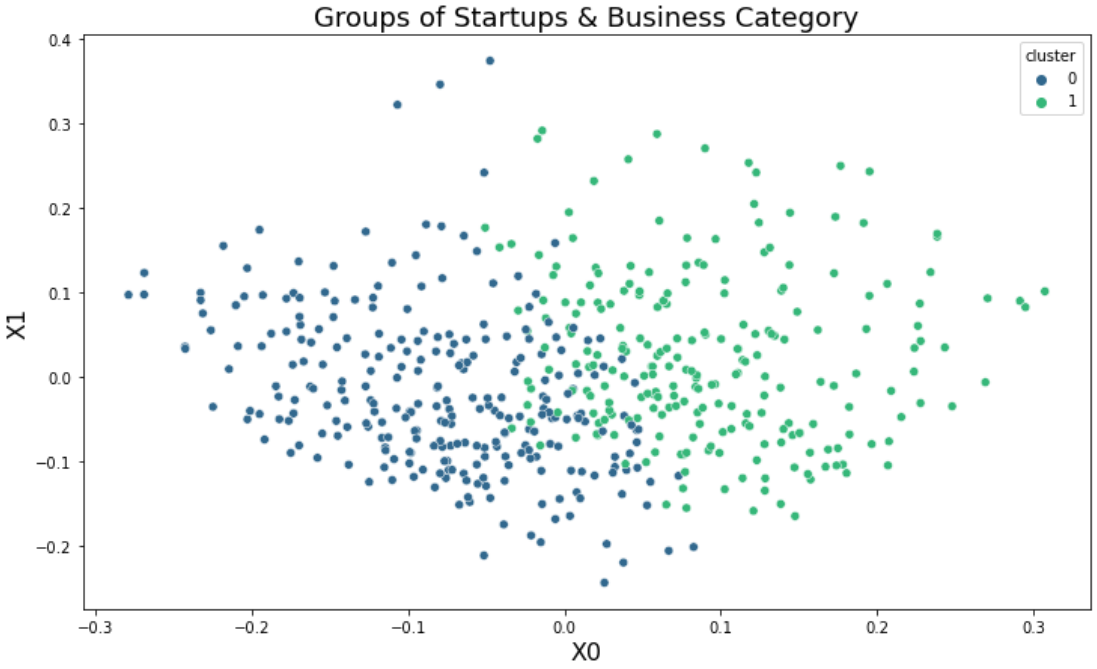



```
the most relevant keywords for each group

Group 0
one,user,share,people,need,company,manager,customer,team,product

Group 1
one,people,get,time,make,like,product,use,user,design
```

Startups & Business Category :




```
the most relevant keywords for each group

Group 0
make,like,company,one,get,thing,time,work,people,team

Group 1
time,one,founder,investor,market,customer,product,startup,business,company
```