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.

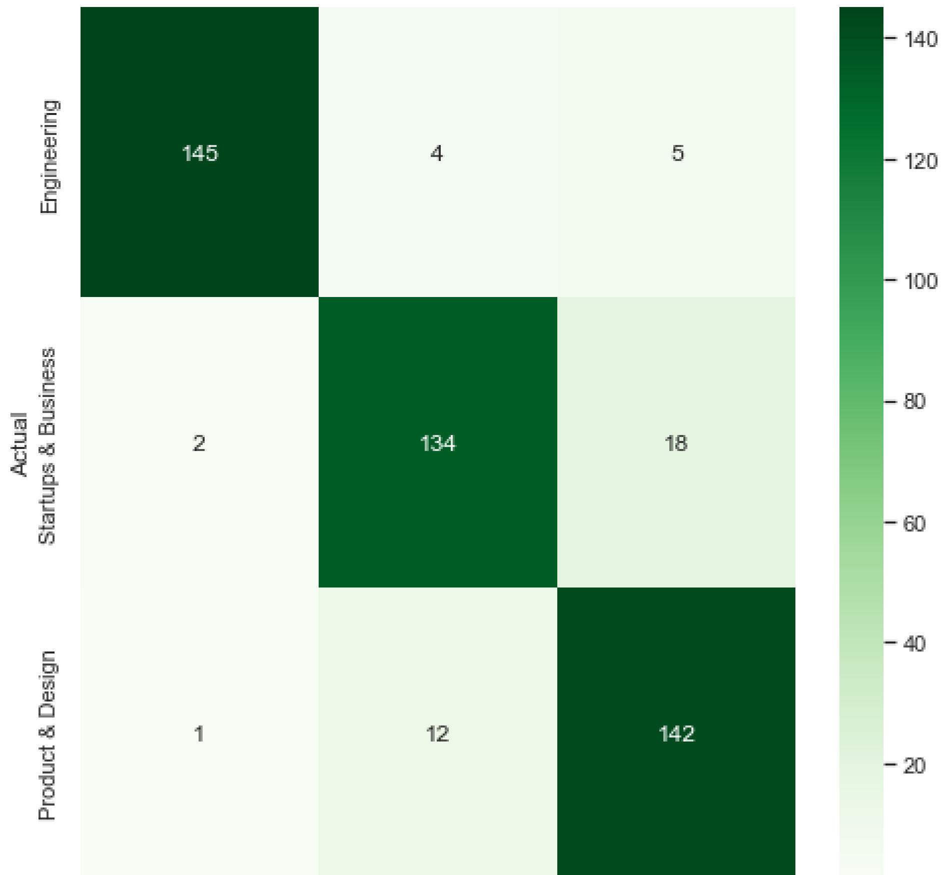
Model Training :

- 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