TEXT CLASSIFICATION USING SCIKIT-LEARN AND NLTK

1. **The Aim:**

   It is to put a set of documents downloaded from a website and manually give the category of the freelancer.

2. **Dataset:**
   - It is a collection of public documents posted by freelancers in their portfolios on the Truelancer website, which was first downloaded and then sorted into ten different domains.
   - The downloaded pdfs were used in the extraction code to extract the text for further classification.
   - The zip file of the Extracted Text was used for this code.
   - Kaggle was used as the medium to import the dataset into google colab using `‘opendatasets’` and the files were loaded using Scikit-Learn.

3. **Feature Extraction:**
   - For running the algorithms, the extracted text files were needed to be converted into numerical feature vectors using Bag of Words.
   - The number of times each word occurs was counted in each document and an integer id was assigned.
   - ‘CountVectorizer’ was used for the same.
   - By doing `‘count_vect.fit_transform(freelancer_data.data)’` it returns a Document-Term matrix - `(274, 8261)`

   TF-IDF: To quantify the words in the document by computing the weight of each word.

4. Naïve Bayes Classifier:

- For classifying the text, Naïve Bayes Classifier was used, which assumes that the presence of a particular feature in a class is not related to any other feature present.
- ‘clf = MultinomialNB().fit(X_train_tfidf, freelancer_data.target)’will train the classifier on the training data.
- After testing the performance of the NB Classifier on the dataset, the accuracy achieved was 54.38%.

5. Support Vector Machine:

- To check for better performance, the SVM algorithm was used for classification.
- After testing the performance of the classifier on the dataset, the accuracy achieved was 98.9%.

6. Grid Search:

- The Grid Search CV library function from sklearn’s model_selection package was used to select the best parameters from the hyperparameters.
- ‘parameters = {'vect ngram_range': [(1, 1), (1, 2)], 'tfidf_use_idf': (True, False), 'clf_alpha': (1e-2, 1e-3)}’ this is to create a list of parameters for tuning to obtain optimal performance that is to use unigrams and bigrams and choosing which is optimal.
- ‘gs_clf_svm = GridSearchCV(text_clf_svm,parameters_svm,n_jobs=-1)’This creates an instance of the grid search that tells to use multiple cores from user machine.
- After checking, the best mean score and best params for both NB and SVM, the accuracy of the classifiers had increased.

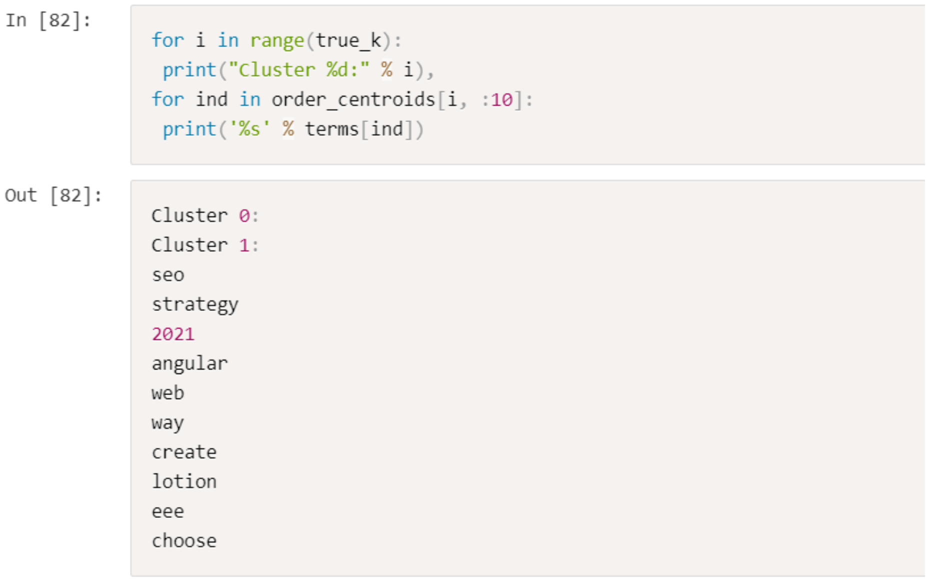
7. Stop Words Removal:

- Stop words are common words that are generally filtered out before processing.
- The same pipeline built for the NB classifier was run.

- NLTK was installed for its various stemmers for reducing the words to root form.
- For the Stemming, An algorithm named Snowball stemmer was used. The accuracy achieved with the stemming was 94.5%, which was a huge improvement in the case of the Naïve Bayes Classifier.

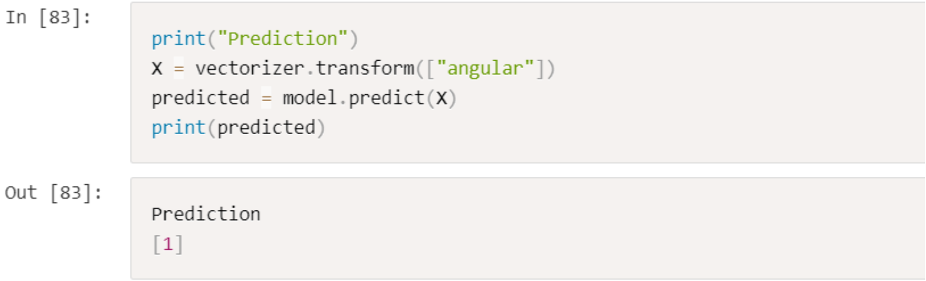
8. K-Means Clustering:

- While importing libraries, TfidfVectorizer was used to evaluate how important a word is in the document.
- Next, a vectorizer using TfidfVectorizer class was created to fit and transform the document.
- K-means clustering algorithm was implemented in the vectorized document.
- K-means clustering algorithm was implemented in the vectorized document.
- Then a code was executed to print the centroids and features into which clusters they belong.



```
In [82]:
for i in range(true_k):
 print("Cluster %d:" % i),
for ind in order_centroids[i, :10]:
 print('%s' % terms[ind])
```

```
Out [82]:
Cluster 0:
Cluster 1:
seo
strategy
2021
angular
web
way
create
lotion
eee
choose
```



```
In [83]:
print("Prediction")
X = vectorizer.transform(["angular"])
predicted = model.predict(X)
print(predicted)
```

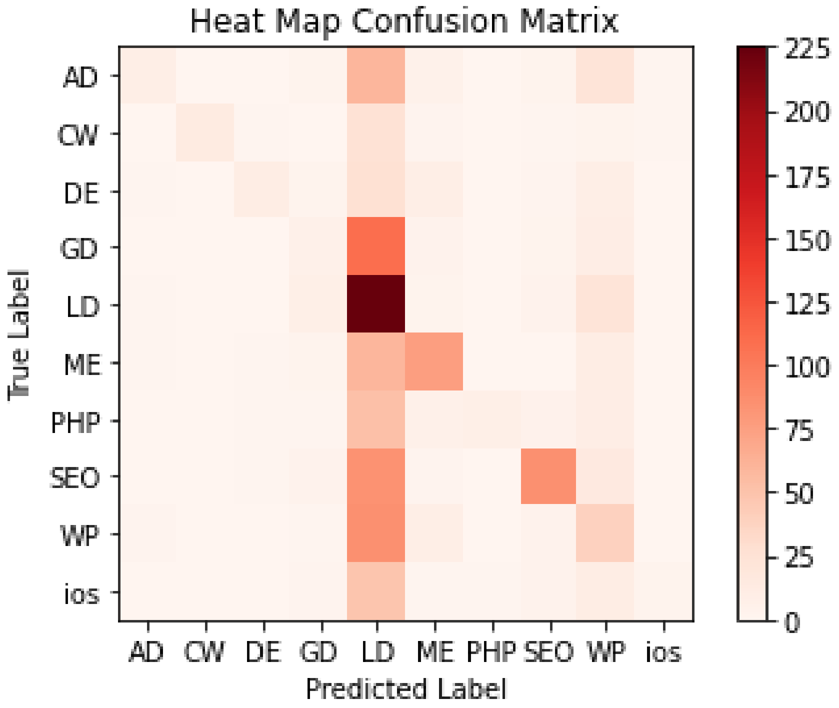
```
Out [83]:
Prediction
[1]
```

- Can also predict a word which tells in which cluster it belongs to.
- Here we get `Prediction [1]` meaning that the word belongs to the cluster so the prediction is correct.

9. Results:

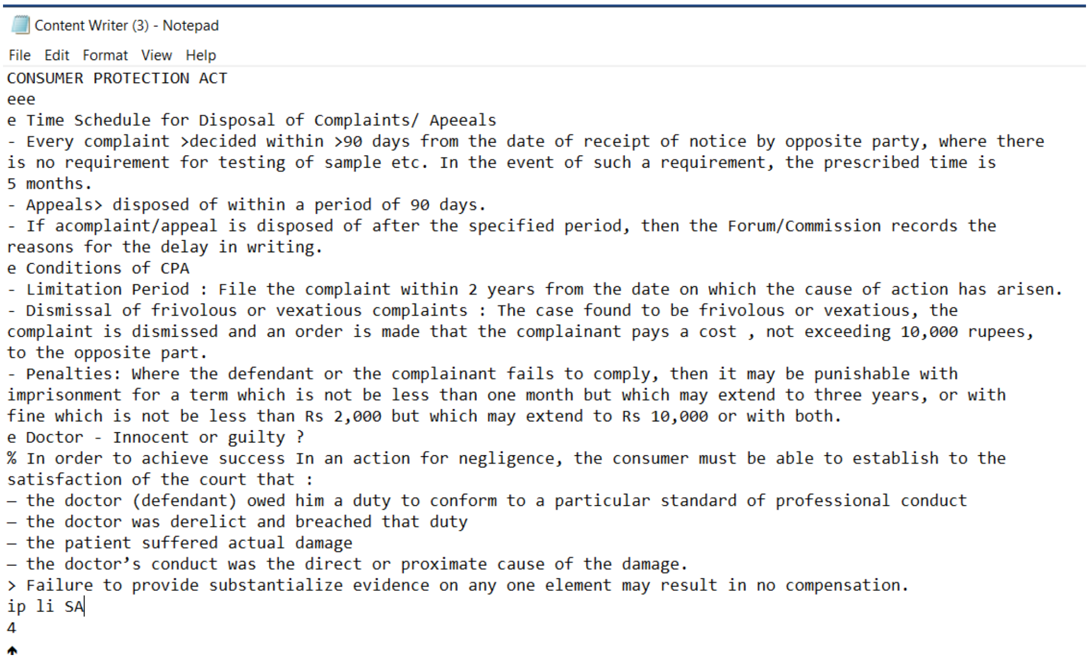
- The input data is the collection of extracted text of all the freelancer pdfs.
- `(274, 8261)`-This output is the dimension of the Document-Term matrix, which is the Bag of Words representation.
- Since Naïve Bayes classifier is a fast and reliable algorithm for large datasets, it produces competitive classification accuracy.
- The base accuracy of `94.5%` seems to make it an effective classifier.
- After labeling the datasets and building the term matrices, a label of the documents in the set was predicted.
- Finally, the Confusion Matrix was visualized.

Following is a Heat Map of the Confusion Matrix to evaluate the quality of the output of the Naïve Bayes Classifier.



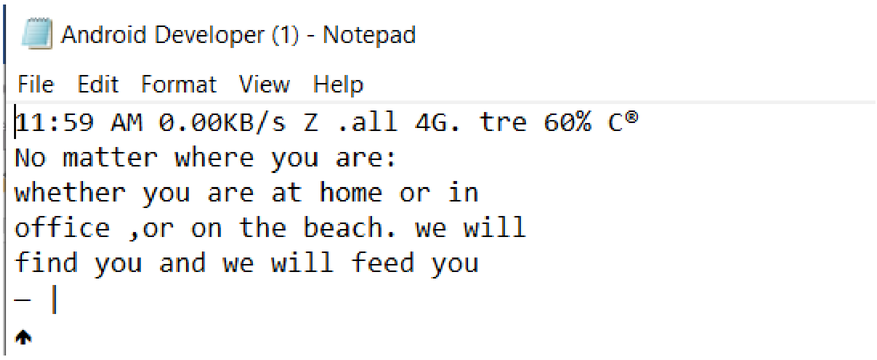
Sample Dataset:

Extracted text from Content Writer Domain:



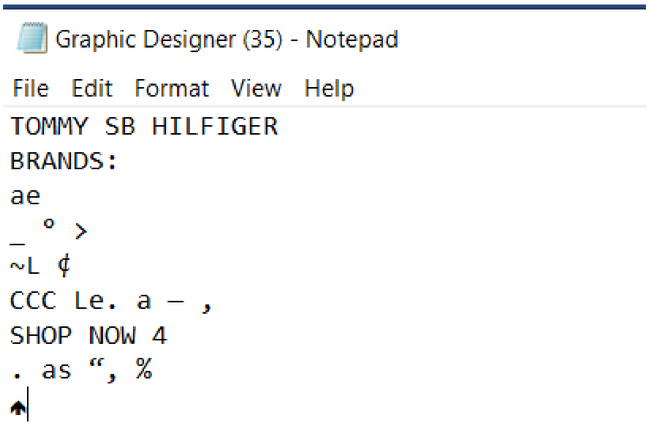
```
CONSUMER PROTECTION ACT
eee
e Time Schedule for Disposal of Complaints/ Apeeals
- Every complaint >decided within >90 days from the date of receipt of notice by opposite party, where there
is no requirement for testing of sample etc. In the event of such a requirement, the prescribed time is
5 months.
- Appeals> disposed of within a period of 90 days.
- If acomplaint/appeal is disposed of after the specified period, then the Forum/Commission records the
reasons for the delay in writing.
e Conditions of CPA
- Limitation Period : File the complaint within 2 years from the date on which the cause of action has arisen.
- Dismissal of frivolous or vexatious complaints : The case found to be frivolous or vexatious, the
complaint is dismissed and an order is made that the complainant pays a cost , not exceeding 10,000 rupees,
to the opposite part.
- Penalties: Where the defendant or the complainant fails to comply, then it may be punishable with
imprisonment for a term which is not be less than one month but which may extend to three years, or with
fine which is not be less than Rs 2,000 but which may extend to Rs 10,000 or with both.
e Doctor - Innocent or guilty ?
% In order to achieve success In an action for negligence, the consumer must be able to establish to the
satisfaction of the court that :
– the doctor (defendant) owed him a duty to conform to a particular standard of professional conduct
– the doctor was derelict and breached that duty
– the patient suffered actual damage
– the doctor’s conduct was the direct or proximate cause of the damage.
> Failure to provide substantialize evidence on any one element may result in no compensation.
ip li SA
4
🠉
```

Extracted text from Android Domain:



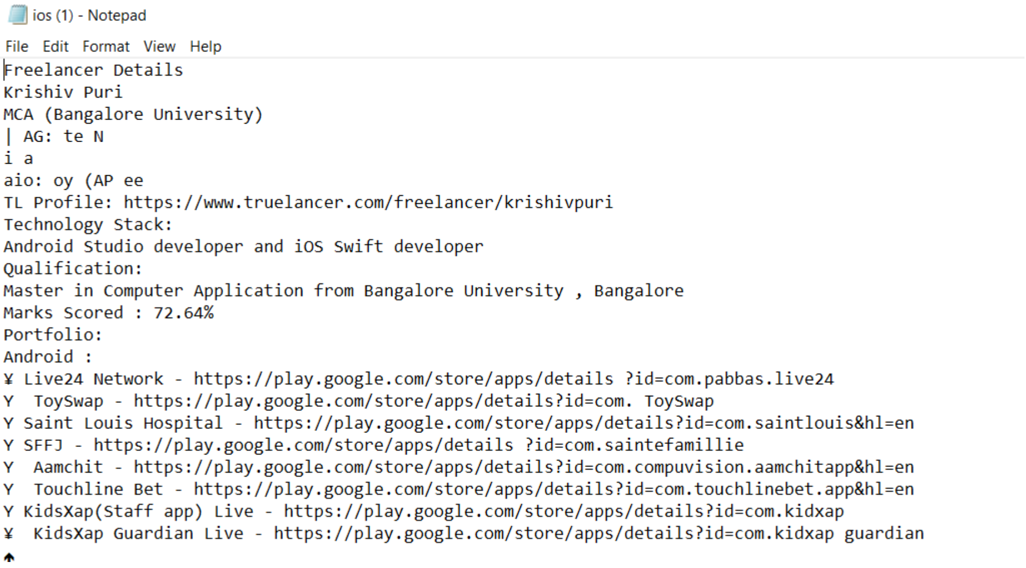
```
11:59 AM 0.00KB/s Z .all 4G. tre 60% C®
No matter where you are:
whether you are at home or in
office ,or on the beach. we will
find you and we will feed you
– |
🠉
```

Extracted text from Graphic Designer Domain:



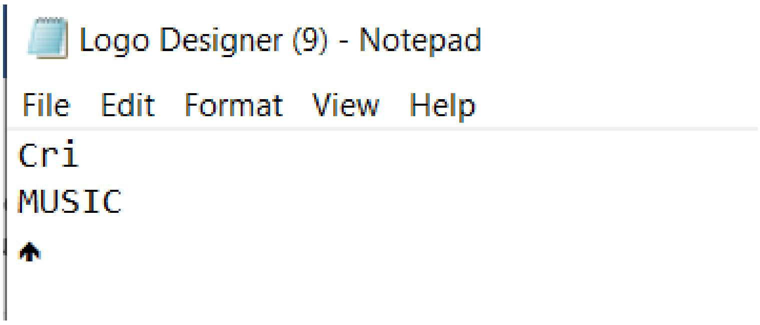
```
TOMMY SB HILFIGER
BRANDS:
ae
_ ° >
~L ¢
CCC Le. a – ,
SHOP NOW 4
. as “, %
🠉
```

Extracted text from ios Developer Domain:



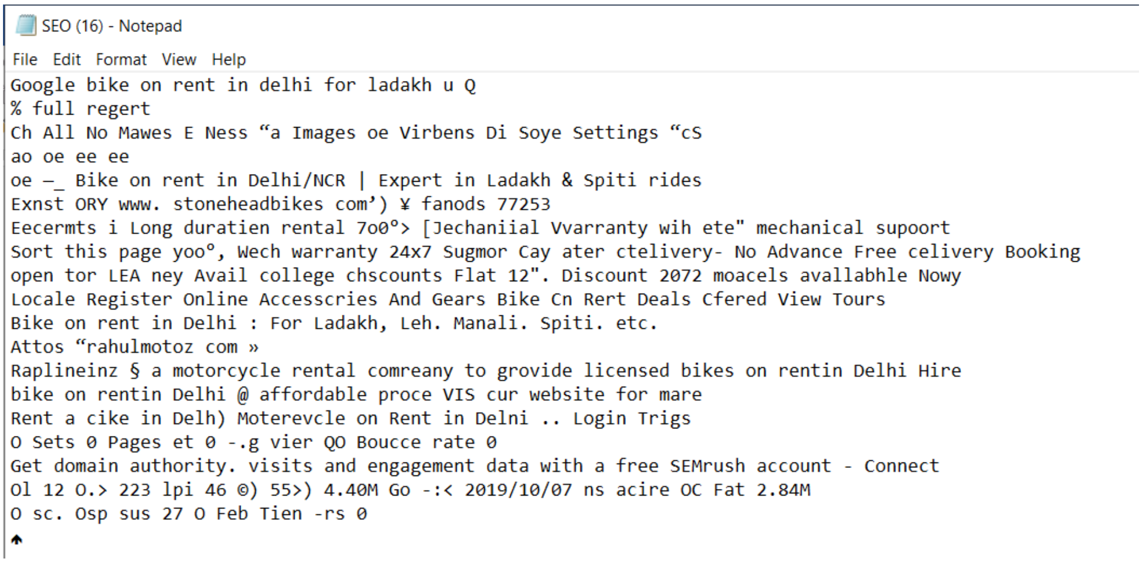
```
Freelancer Details
Krishiv Puri
MCA (Bangalore University)
| AG: te N
i a
aio: oy (AP ee
TL Profile: https://www.truelancer.com/freelancer/krishivpuri
Technology Stack:
Android Studio developer and iOS Swift developer
Qualification:
Master in Computer Application from Bangalore University , Bangalore
Marks Scored : 72.64%
Portfolio:
Android :
¥ Live24 Network - https://play.google.com/store/apps/details ?id=com.pabbas.live24
Y  ToySwap - https://play.google.com/store/apps/details?id=com. ToySwap
Y Saint Louis Hospital - https://play.google.com/store/apps/details?id=com.saintlouis&hl=en
Y SFFJ - https://play.google.com/store/apps/details ?id=com.saintefamillie
Y  Aamchit - https://play.google.com/store/apps/details?id=com.compuvision.aamchitapp&hl=en
Y  Touchline Bet - https://play.google.com/store/apps/details?id=com.touchlinebet.app&hl=en
Y KidsXap(Staff app) Live - https://play.google.com/store/apps/details?id=com.kidxap
¥  KidsXap Guardian Live - https://play.google.com/store/apps/details?id=com.kidxap guardian
```

Extracted text from Logo Designer Domain:



```
Cri
MUSIC
```

Extracted text from SEO Domain:



```
Google bike on rent in delhi for ladakh u Q
% full regert
Ch All No Mawes E Ness “a Images oe Virbens Di Soye Settings “cS
ao oe ee ee
oe —_ Bike on rent in Delhi/NCR | Expert in Ladakh & Spiti rides
Exnst ORY www. stoneheadbikes com’) ¥ fanods 77253
Eecermts i Long duratien rental 7o0°> [Jechaniial Vvarranty wih ete" mechanical supoort
Sort this page yoo°, Wech warranty 24x7 Sugmor Cay ater ctelivery- No Advance Free celivery Booking
open tor LEA ney Avail college chscounts Flat 12". Discount 2072 moacels avallabhle Nowy
Locale Register Online Accesscries And Gears Bike Cn Rert Deals Cfered View Tours
Bike on rent in Delhi : For Ladakh, Leh. Manali. Spiti. etc.
Attos “rahulmotoz com »
Raplineinz § a motorcycle rental comreany to grovide licensed bikes on rentin Delhi Hire
bike on rentin Delhi @ affordable proce VIS cur website for mare
Rent a cike in Delh) Moterevcle on Rent in Delni .. Login Trigs
O Sets 0 Pages et 0 -.g vier QO Boucce rate 0
Get domain authority. visits and engagement data with a free SEMrush account - Connect
Ol 12 O.> 223 lpi 46 ©) 55>) 4.40M Go -:< 2019/10/07 ns acire OC Fat 2.84M
O sc. Osp sus 27 O Feb Tien -rs 0
```

10. Conclusion:

- Performed the Classification of the Extracted text using Naïve Bayes and SVM Classifiers.
- Performed Grid search for tuning the performance and used NLTK stemming approach.

11. References:

1. https://www.kaggle.com/ankitakokes/extracted-text
2. http://www.nltk.org/
3. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.Multinomial N B.html
4. https://scikit-learn.org/stable/modules/svm.html
5. https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes
6. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.C o untVectorizer.html
7. https://github.com/javedsha/text-classification
8. https://towardsdatascience.com/machine-learning-nlp-text- classification-using-scikit-learn-python-and-nltk-c52b92a7c73a
9. https://github.com/joaorafaelm/text-classification-python