identify and classify documents by topic

Problem description

the idea: identify and classify documents by topic
we will use dataset scrapped from Makeup site and make classification on it
This dataset contains makeup products from the Heruko App Makeup API,
It has different prices, descriptions, names, price_sign, currency,
product_types like(lip_liner, lipstick, foundation, eyeliner, eyeshadow and so on),
brands like(colourpop, boosh, deciem, zorah biocosmetiques and so on),
categories like(like lipsticks, liquids, powder, lip_gloss, pencil and so on), and
tags (like Gluten Free, Vegan, silicone free, oil free, alcohol free, cruelty free, water
free and so on).

** dataset link: https://github.com/aniass/Extracting-data-using-API

Model design

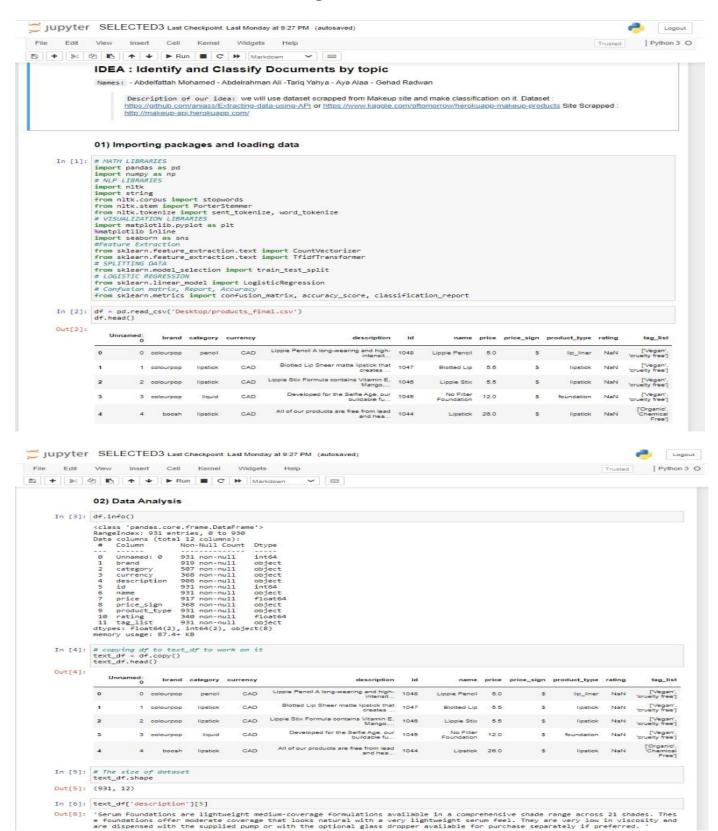
steps	
01) Importing packages	-Import all the packages needed
and loading data	-Reading the data
02) Data Analysis	-Check missing values
	-Data type changed to string
	-making the data unique
	-Grouping data with similar names
	,After running we will have five
	group of classes:
	(Eye makeup-Lipstick-Foundation-
	Contour-Nail polish)
	-Mapping data to numeric data for
	learning
03) Text Pre-processing	-remove punctuations
	-remove stopwords
	-Stemming
	-COUNT VECTORIZER for ENCODING
	-Extract TF-idf
04) Splitting Data	-Splitting the dataset into train and
	test data
05) Build the models	-we used (logistic Regression-SVM-
	Random Forest)

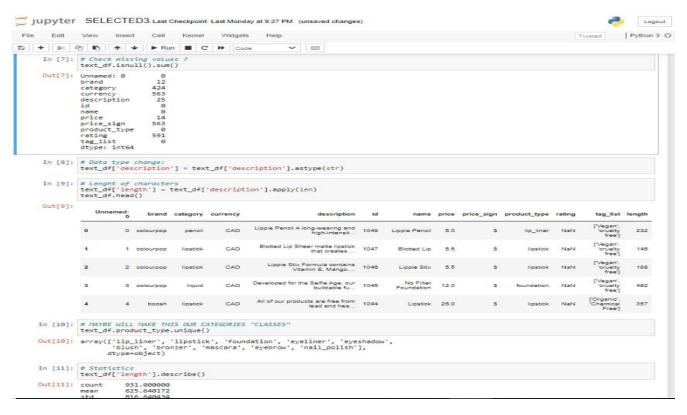
Experimental results

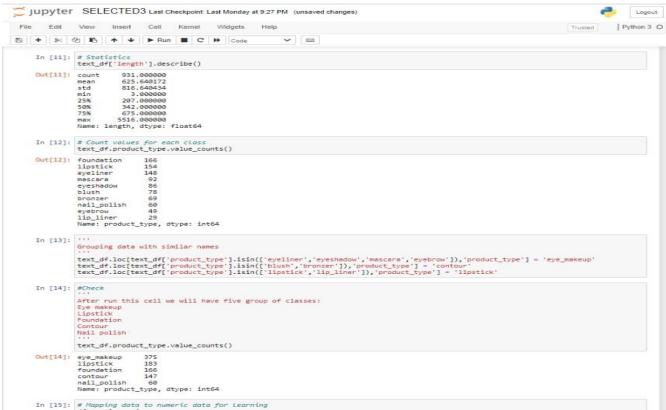
We used three classifiers (logistic regression-svm-random forest)

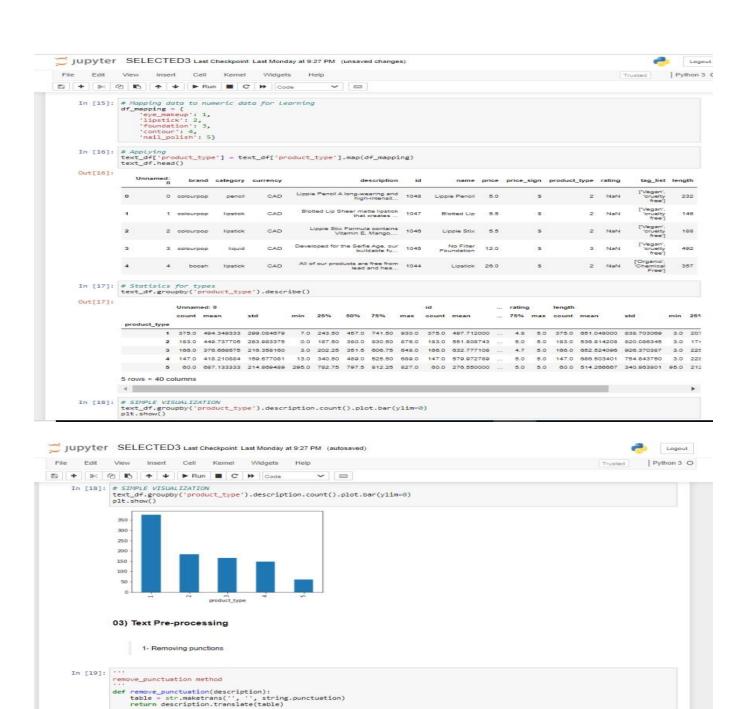
Classifier	Accuracy
logistic Regression	0.8669527896995708
SVM	0.9227467811158798
Random Forest	0.9012875536480687

----Pictures of the code running----









description id name price price_sign product_type rating

tag_list length

232

188

2 NaN

2 NaN

3 NaN

In [20]: # Applying
 text_df['description'] = text_df['description'].apply(remove_punctuation)
 text_df.head()

0 colourpop pencil CAD Lippie Pencil A longwearing and high intensity ... 1048 Lippie Pencil 5.0 \$

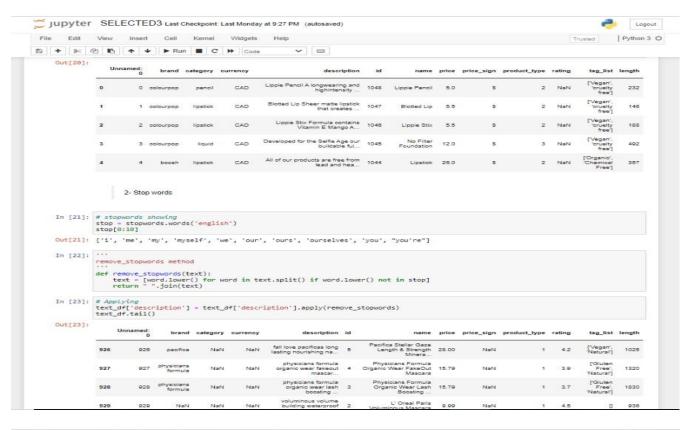
2 colourpop lipstick CAD Lipple Stix Formula contains Vitamin E Mango A... 1046 Lipple Stix 5.5

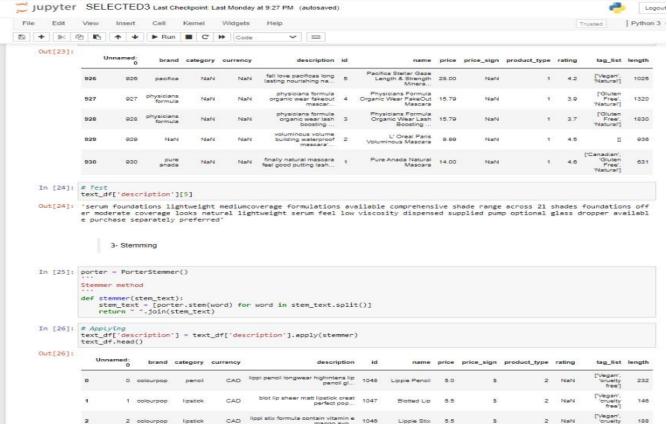
3 colourpop liquid CAD Developed for the Selfle Age our buildable ful... 1045 Foundation 12.0

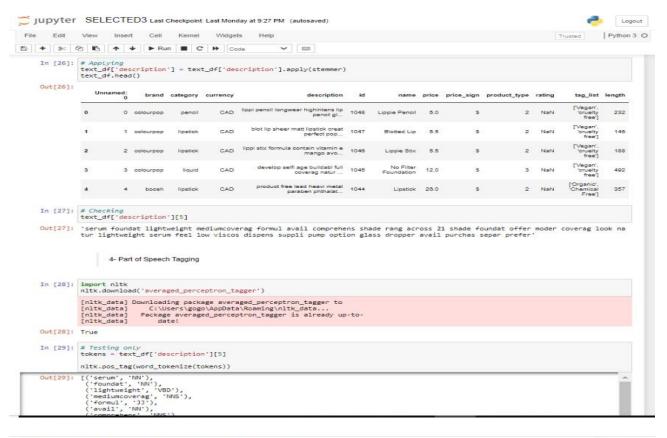
CAD Blotted Lip Sheer matte lipstick that creates ... 1047 Blotted Lip 5.5

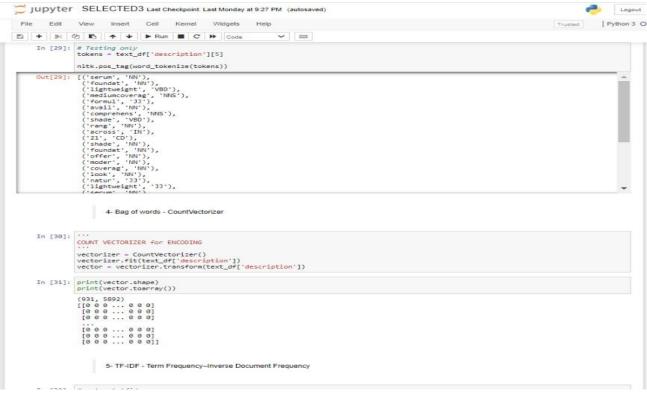
Unnamed: brand category currency

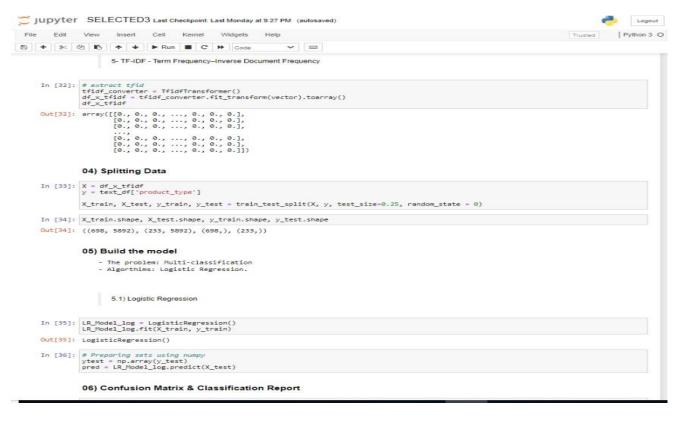
Out[20]:

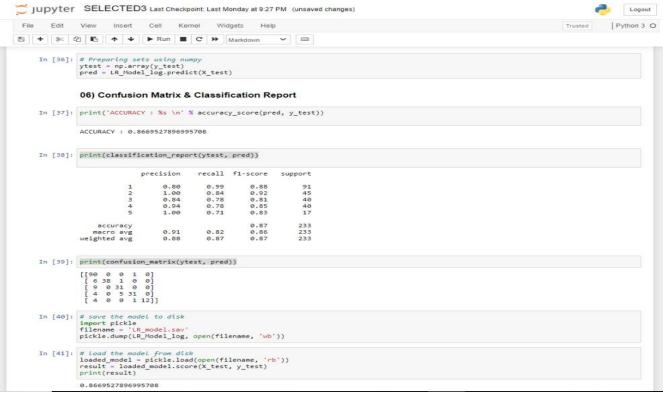


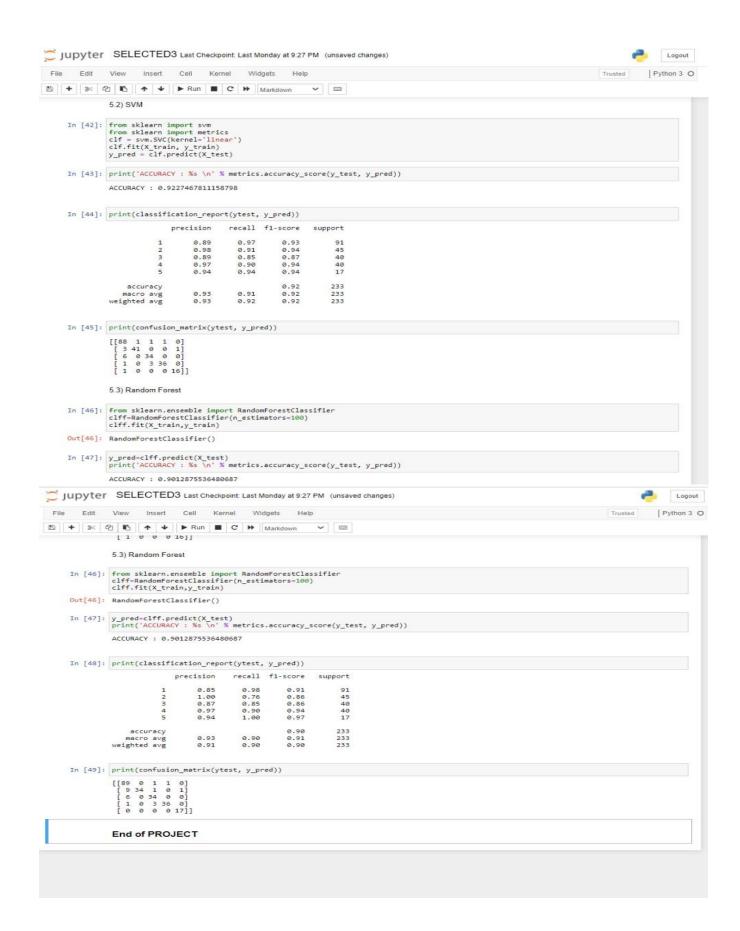












Model performance

the best model performance is SVM with accuracy 92%