Module 4: Text Classification 1

Case Study

edureka!

edureka!

© Brain4ce Education Solutions Pvt. Ltd.

Case Study

The data for this assignment is taken from Kaggle.
(https://www.kaggle.com/zynicide/wine-reviews)

This is a subset of the original data which contains information such as Country , Description , Points and Designation of certain types of Wines . You have to use the perform the following functions on this file

- Read the CSV file "Wine.csv"
- Using Pre-process File which you created in module 2 case study 1, call the 'Refine' function and get the pre-processed text for each 'description' in the csv file. Store it in a column named "Refined-Description"
- Using 'CountVectorization' function from 'Vectorization' python file created in case study 1of this module, vectorise all the rows in 'Refined-Description' column which you created in the above step. Store them in a column named "CountVectorizer"
- Using 'TF-IDFVectorization' function from 'Vectorization' python file created in case study 1 of this module, vectorise all the rows in 'Refined-Description' column and store the results in 'TF-IDF Vectorizer Column"
- Save changes to the CSV file.

NOTE: We have created 3 Python Files till now

1. *PreProcess:* Has functions to tokenize, remove stop words & lemmatize string
2. *Corpus*: Takes multiples strings/documents and returns a corpus
3. *Vectorization:* Has functions for presence absence, count and tf-idf vectorization

These will be helpful in upcoming case studies and projects.