Module 5: Text Classification 2

Case Study

Case Study

Epinions.com is a website where people can post reviews of products and services. It covers a wide variety of topics. For this case study, we downloaded a set of 600 posts about digital cameras and cars and saved as "Eopinions.csv".

The dataset has 2 columns: 'class' and 'text'

	class	text
0	Auto	I have recently purchased a J30T with moderat...
1	Camera	This camera is perfect for anyone who wants t...
2	Auto	2000 Hyundai Elantra Wagon if you can find ...
3	Camera	I bought this product because I need instant ...
4	Camera	Before I begin my objective review I should ...
5	Camera	I bought the Minolta Dimage 2300 almost a yea...
6	Camera	Life can t get any better than this After ow...
7	Camera	I ordered this from a discount yahoo shop wit...
8	Camera	I bought this camera on ebay The FD71 model ...
9	Auto	I have owned my Buick since 53000 km and I am...

These are the tasks which you have to perform:

- Read the file as a pandas data-frame.
- Perform Label Encoding on 'class' column.
- Plot a bar graph to compare the frequencies of both the classes.
- Preprocess the 'text' column
- Vectorize the text using CountVectorizer
- Split the dataset into 2 parts namely "train.csv" and "test.csv" having 80% and 20% of the data respectively from the original data. These are your Train and Test Data. Make sure train and test data are having same proportion of data points as the original data
- Train your machine learning algorithm for classification and prepare a model (you can choose any appropriate algorithm of your choice)
- Now test the model on the Test data and evaluate the Performance by providing Confusion Matrix for your model.
- Plot ROC Curve.