Project:

Twitter Sentiment Classification

Learning Objectives

To implement the techniques learnt as part of the course.

- Text Based Exploratory data analysis
- Pre-processing of text data.
- Vectorization of text data
- Building Classifier
- Tune the Classifier
- Evaluate the Classifier

Data Description

- A sentiment analysis job about the problems of each major U.S. airline.

- Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

- It contains whether the sentiment of the tweets in this set is positive, neutral, or negative for six US airlines:

Data Description

Dataset:

- The project is from a dataset from Kaggle.
- Link to the Kaggle project site:

https://www.kaggle.com/crowdflower/twitter-airline-sentiment

- The dataset can be downloaded from the above Kaggle website or from Olympus.

Steps to follow

Import the libraries, load dataset, print shape of data, data description.

- Do Exploratory data analysis
- Understand of data-columns:
  - Drop all other columns except “text” and “airline_sentiment”
  - Check the shape of data
  - Print first 5 rows of data.
- Text pre-processing: Data preparation.
  - Html tag removal
  - Tokenization.
  - Remove the numbers.
  - Removal of Special Characters and Punctuations.

Steps to follow

- Text Pre-processing (continued...)
  - Conversion to lowercase
  - Lemmatize or stemming
  - Join the words in the list to convert back to text string in the dataframe. (So that each row contains the data in text format.)
  - Print first 5 rows of data after pre-processing.

- Vectorization:
  - Use countvectorizer
  - Use TfidfVectorizer

Steps to follow

- Fit, tune and evaluate the model using both types of vectorization.
- Print the top 40 features and plot their word cloud using both type of vectorization.
- Summarize your understanding of the application of Various Pre-processing and Vectorization and performance of your model on this dataset.

Questions?

Thank You.
Happy Learning!