NLP => Spam detection

Sms spam detection :-

#1 App.py (python run file)
Pickle :- serializing and de-se-. (converts obj into byte)
render_template request :- used for Jenrate output from a template file based on jinja2 engine
-> 'rb' => Read binary.

#2 Import Pandas (for Analysis and data structure, array)
Pickle
(1) Importing libraries, nltk, re (regular expression)
(2) load dataset,
(3) download (stop words)
(4) Porter Stemmer (reducing suffix and prefix, also common words)
(5) Cleaning the message.
-> (i) Clean special characters (#, @, etc)
(ii) Convert message into lower case
(iii) Tokenizing (split full message into words)
(iv) Remove stop words
(v) stemming the words (grew, growing = grow)
(vi) Joining the non-stemmed words grew
(vii) Build a Corpus of message (Collection of words).
(6) Create the model.
sklearn, feature extraction (get only raw data, get only useful data)
CountVectorizer (Convert text into vector.
max_features (used for build Vocabulary. only consider top features according to corpus.
cv.fit_transform(corpus).toarray (Convert text into vector)

(7) Extracting dependent Variable from data set.

→ get_dummies (give 0 or 1 to label according to labels)

→ iloc (If you want to use specific rows/cols from data, used loc. and iloc)

(8) Creating a pickle file for the Count Vectorizer

pickle.dump(cv, open('cv.pkl', 'wb'))

(9) model building :

from sklearn.model_selection Import train_test_split

test_size = 0.20 (train = 80%, test = 20%.)

to training set.

(10) naive Bayes

Import multinomialNB

Classifier = multinomialNB(alpha=0.3) naive bayes classifier = used for Predict the Probability of a various Attributes.)

↑ α = laplace smoothing

.fit <= fit model in training

(11) Create Pickle file for multinomial model.

filename = 'spam.pkl', pickle.dump(classifier, open(filename, 'wb'))