AIT 726 Airline Tweets Dataset Processing

Programming Hints and Tips*

Importing data

As dataset is given as documents with each document in separate folder. To import such a type of documents use **OS module.**

Steps:

- Use chdir(path) method of os module with path of dataset folder as its input. It basically moves to child directory with the path
- Use glob() method of glob module with *.txt method and store result in one variable. This basically access all text files in that folder and iterating through thevariable which is used to store the .txt files we can import the data.

Preprocessing

The imported tweets are stored in a list which are then sent as input to the preprocessing method. Which basically removes HTML tags using BeautifulSoup, translates emojis to text using emoji.demojize() method, replaces all unnecessary punctuations with whitespaces etc.

After preprocessing tweets send them along with one flag variable into stemmer method wherewe use either snowball or porter stemmer to stem the tweets. Here the flag variable whether ornot to stem tweets.

NOTE: when one tries to lower case all words other than acronyms there might be some exceptions like MY which may be treated as Acronym. Model tries to interpret “MY” as different from “my”.
Also, it’s better to split punctuations as they convey more weight. For example: model tries to learn that “!” is used as sign of positive and when you tokenize happy!!! To happy and “!”,”!”,”I”. The model understands that there are 3 exclamations, and it votes more towards positive ness. On the other hand, if you leave it as “!!!” then model thinks that “!!!” as new word which Is different from “! “ And finally, one can remove links and html tags as links doesn’t add much context to our data. But you can still explore by with and without removing links and just removing tags.

Creating count vectorizer

The preprocessed tweets are sent as input.

- Create a dictionary say counts.
- Counts[1]=defaultdict(int).
- Counts[0]=defaultdict(int)
- Keep words as keys and increase count of word for each time of its occurrence in thedataset

This way of using default dictionary. We can store count of words in positive tweets andnegative tweets separately**.**

** Prepared by GTA Rajeev Kashetti and Professor Liao.*

* Prepared by GTA Rajeev Kashetti and Professor Liao.