Natural Language Processing – Assignment 4

Handed Out:	1st September 2021
Due Date	11th September 2021, midnight.
Expected deliverables	One compressed electronic file containing the reports and code specified below.
Method of Submission:	Online via Moodle (see below).

Assignment Description

Organizations are eager to gather the views of their main stakeholders (referred to as customers hereinafter) via online forums. One of the common ways to get a 'birds eye' view of this is to use sentiment analysis on the feedback they receive. More importantly, they are particularly interested in obtaining any suggestions that their customers have on how to improve their offerings. The task below is to provide such organizations a way to do this and to do it so that suggestions on the different aspects of their service offerings are categorized in order to give appropriate priority to the various suggestions received.

You are required to produce Python 3 code in a Jupyter Notebook to do the following.

(a) Use the *hotel reviews* subset (only) of the suggestion mining dataset provided at https://github.com/sapna13/Suggestion-Mining-Datasets. Read the csv data files (training and testing) into a single merged Pandas *dataframe* and clean the data in appropriate ways. Print the number of posts which are suggestions and the number that are non-suggestions (opinions) in order to gauge the dimensions of the dataset. Also print the number of total words in this dataset and the number of unique words in it.

(b) Other ways of reducing 'noise' in the dataset are to (i) remove the so called *stopwords* and (ii) to *stem* or *lemmatize* the rest of the words. Print the number of total words and the number of unique words after each of the above two steps for this dataset. Also print the maximum and minimum document sizes (in words) of the posts after each step.

(c) Create *bag-of-words* and *TF-IDF* representations of the posts in the dataset above[1] and use two relevant *supervised learning* algorithms to classify future posts as suggestions or non-suggestions. Print the *confusion matrices* of the four (04) resulting combinations for a held-out (test) dataset.

[1] Using the *CountVectorizer* and *TFIDFVectorizer* in *scikit-learn*.

(d) Suggest any strategies you may use to *improve* the performance of the above classifier (apart from using deep learning). Implement your suggestions as improvements to the above models and print the confusion matrix of the best representation and model you get.

(e) In general, suggestions can be categorized into particular *aspects* of the service provided by the hotel. Propose an *unsupervised* approach to categorize the suggestions (ignoring non-suggestions) into the different aspects of the service offered. How could an *optimal* set of categories be determined? Implement your solution and print the optimal number of categories your model finds. Visualize your categories to see if they are reasonable.

(f) Assume that the hotel gives you the *attached* 'dirty' set of *manually annotated* suggestions[2] and evaluate the performance of your best model with respect to this *ground truth* data. Print appropriate metrics to evaluate your model.

(g) Discuss (without implementing) any issues with the data and models used, and any improvements you would suggest to make the overall model more useful and/or perform better. A reflection on the coursework task as a whole is expected.

Submission

You need to formulate solutions for each of parts (a) through (g) above, clearly explaining your Python code and specifying the outputs produced by the code for the dataset given in a *Jupyter Notebook* named *Solution_IDNumber.ipynb* based on the template given[3]. For each such part, a descriptive summary with an interpretation should be given for the output obtained after each executable *cell*. Your Jupyter Notebook should be submitted to the VLE *as a single compressed (.zip or .rar) file* with the above naming (e.g. *Solution_17001123.ipynb.zip*).

[2] In the file named 'Test Data'.

[3] The *IDNumber* part of the filename should be replaced with your *UCSC index* number.