

Machine Learning (CSE-475)

Due date: 11/11/2018

Assignment 2: Text Classification with Naïve Bayes

In this assignment, you will classify customer reviews of a restaurant using Naïve Bayes classifier. Write Python code to solve the above task.

Preliminaries:

Read and understand the lecture slide of Naive Bayes classifier and learn the technical details of the implementation. If you wish, you may also have a look at some academic papers:

- **Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan:** [*Thumbs up? Sentiment Classification using Machine Learning Techniques*](#). In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).
- **Thorsten Joachims:** [*Text Categorization with Support Vector Machines: Learning with Many Relevant Features*](#). Technical report from the University of Dortmund, 1997.
- **John Blitzer, Mark Dredze, and Fernando Pereira:** [*Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification*](#). In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007).

Data Set and input/output format:

Download the Restaurant_Review.tsv file. This file has two columns including “Review sentences” and their sentiment category label (1 or 0) in “Liked” column. Sentiment 1 represents positive and 0 represents negative review from a customer. There are 1000 customer reviews are presented in the text file. The data has been presented so that there is one review per line, and the texts have been tokenized and normalized. Here is an example of a line:

```
Wow... Loved this place. 1
```

80% of the total data will be used as training and 20% of them use for testing.

Output: You need to find out the accuracy and confusion matrix for Naïve Bayes Classifier.