Machine Learning Assignment 1: Classification, natural language
Machine Learning, PSAM 5020, Spring 2018

First iteration due: Monday, March 5 at 5:00pm
Final version due: Monday, May 7 at 5:00pm

Data: Amazon Fine Food Reviews The Amazon Fine Food Reviews dataset consists of 455,000 food reviews Amazon users left up to October 2012.

Column	Description
Id	
ProductId	unique identifier for the product
UserId	unique identifier for the user
ProfileName	
HelpfulnessNumerator	number of users who found the review helpful
HelpfulnessDenominator	number of users who indicated whether they found the review helpful
Score	rating between 1 and 5
Time	timestamp for the review
Summary	brief summary of the review
Text	text of the review
helpScore	HelpfulnessNumerator / HelpfulnessDenominator
helpful	boolean; =True if review is “helpful”

The data is available at the course Canvas page: `Amazon.csv` in “files.” The dataset is a subset of a publicly available dataset. **Important: do not use the publicly available dataset in training or assessing your model.** Data source reference: J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.

Classification Task
You are going to train a model to best predict `helpful`, a boolean indicator that a review was deemed helpful by other shoppers. This was defined as:

```
(data.HelpfulnessDenominator > 3) & (data.helpScore >= 0.9)
```

For this assignment, you must use only the data available in `Amazon.csv`; do not augment the data with other sources. However, the features will need extensive transformations before they will be useful for training the model. A resource for text feature extraction: http://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

You are encouraged to create features beyond those available in the feature extraction documentation. Possibilities include the length of the review, its positivity or negativity, its grammar, etc. https://www.amazon.com/dp/B077J9JFC5

Do not use any of the following variables in `Amazon.csv` (or variations of them) as features:

- helpful
- helpScore
- HelpfulnessNumerator
- HelpfulnessDenominator

Your work will be assessed on:

- how accurately your model classifies on a test set
- how well your model generalizes
- the organization and documentation of your Jupyter Notebooks
- communication of your work in class reflections and final presentations
- model improvement over the semester

Submission instructions:

- Each submission will consist of one zip file, uploaded to Canvas. The Zip file must contain one Python Notebook (.ipynb) and multiple pickle files (.pkl), and *optional* the my_measures module (.py). It should NOT include any of your raw data files (.csv). (Note: The pickle files and NumPy files would have been generated as part of training the data.)
- The Python Notebook file should do the following:
    - Read in the CSV file that contains your raw (unaltered) **TEST** data. Again, you do not need to include this CSV file in your submission.
    - Go through all of the same steps involved in constructing your **`x`** matrix and **`y`** vector that you completed on the **TRAINING** data. One important difference: for all class instances that involved data transformations and/or model fits, you should be loading the "pickled" instance and only using methods that transform or predict. Very important: in this notebook, you should not be using any methods that fit (including methods that also do other things, like "fit_transform." Any use of a method that includes the word "fit" will automatically assigned a value of 0 for all performance measures.
    - With **`x`** and **`y`**, predict **`y`** ***using the pickled model fit*** and calculate and display the relevant performance measures. The notebook should include only one model: the one that is performing best for you. If you include more than one model, I will only grade the first one.

- The notebook should include all your output and results.