Applied Data Analytics:
Assignment #1
Assigned readings
sk-learn tutorial at http://scikit-learn.org/stable/tutorial/basic/tutorial.html#
SKLearn Feature-extraction documentation: http://scikit-learn.org/stable/modules/feature_extraction.html
Pandas tutorial at http://pandas.pydata.org/pandas-docs/version/0.18.1/tutorials.html
NLTK book, chapter 3: http://www.nltk.org/book/ch03.html

Prior to completing this homework, download and install sklearn, matplotlib, & numpy to your machine. Provide screenshots showing successful imports of these packages.

1) (10 points) Provide a short definition of the following terms **IN YOUR OWN WORDS (no copy-paste)**. For each term, if relevant, indicate whether it is a supervised learning task or an unsupervised learning task. **Provide an example of each of these tasks:**
   a. Classification
   b. Clustering
   c. Density estimation
   d. Regression
   e. Samples
   f. Features
   g. Multivariate data
   h. Training set
   i. Test set
   j. Holdout set

2) (10 points) Implement the tutorial found at http://scikit-learn.org/stable/tutorial/basic/tutorial.html#introduction

   For each line of code, write a short sentence describing the meaning of that code (**in your own words**). Write a short paragraph describing the overall function of the code. Treat the Support Vector Classifier as a "black box" – don't worry about how it works (that's coming later).

   Implement the tutorial here: http://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html#example-classification-plot-digits-classification-py

   For each line of code, write a short sentence describing the meaning of that code (**in your own words**). Write a short paragraph describing the overall function of the code. As above, treat the Support Vector Classifier as a "black box" – don't worry about how it works (that's coming later).

3) (10 points) Modify the digits tutorial for the Iris dataset (use 125 randomly selected samples as training data and the last 25 samples as test data).

4) (10 points) Load the "Boston Housing" dataset from SKLearn. Save it to a Pickle (.pkl) file and submit the file with your assignment.
    a. How many samples does this dataset have?
    b. How many features does this dataset have?
    c. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?
    d. Justify, in writing, your answers to part c. If a feature can be both categorical or continuous, present an argument for why your assignment is correct.
    e. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

5) (10 points) Load the airports only (airports.dat) dataset from http://openflights.org/data.html into SKLearn using the dictionary vectorizer function.
    a. First, use the read_csv function in pandas to load the data into a dataframe. Next, use the to_dict function to save the data as a dictionary (HINT: transpose the data and make sure to only keep the values)
    b. Use the sklearn DictVectorizer to load the data
    c. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.
    d. How many samples does this dataset have?
    e. How many features does this dataset have?
    f. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?
    g. Justify, in writing, your answers to part iii
    h. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

6) (10 points) Load a non-JSON dataset of your choice from data.gov, http://catalog.data.gov/dataset into SKLearn. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.
    a. Spend some time exploring this dataset, and brainstorm a question that you can use it to answer.
    b. First, use the read_csv function in pandas to load the data into a dataframe. Next, use the to_dict function to save the data as a dictionary (HINT: transpose the data and make sure to only keep the values)
    c. Use the sklearn DictVectorizer to load the data
    d. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.
    e. How many samples does this dataset have?
    f. How many features does this dataset have?
    g. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?
    h. Justify, in writing, you answers to part iii
    i. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)
    j. Propose a plan to use this dataset to answer your question in part a.

7) (10 points) Choose a (different) JSON dataset for #6.

For the following problems, download and install nltk, urllib2, bs4, feedparser, pypdf, and tweepy to your machine. Provide screenshots showing successful imports of these packages.

8) (30 points) Download text from each of the websites listed below. For each data source,

   a. list the top most frequently occurring bigrams (HINT: use the nltk collocations() function with 30 as an input).
   b. Split the text into documents
   c. Generate two term-document matrices from this dataset (one where each unigram is a token, and one where bigrams can also be tokens). Import the matrix into sklearn and save it as a PKL file. For each text file, indicate how many terms, documents, and unigram tokens are in the corpus

      i. George Washington's Masonic Correspondence (UTF8) from Project Gutenberg. Treat each paragraph as one "document"
      ii. FDA Circulatory System's Devices Panel Advisory Panel Meeting of June 23, 2005.
          Hint: use Beautiful Soup. One document is the end of a speaker's statement
      iii. Nate Silver's Sports RSS feed http://fivethirtyeight.com/sports/feed/ One document is an article
      iv. NASA's Systems Engineering Handbook http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20080008301.pdf -- hint: use pypdf. One document is denoted by a carriage return
      v. A sample of 10,000 tweets on a search string of your choice – hint: use tweepy or SFM. One document is a tweet. (If you are having trouble getting your own Twitter API credentials, pick a publicly available dataset that uses full JSON encoding).
          1. Extra credit: Redo your analysis on a sample of tweets that does not include retweets and tweets with URLs
      vi. Download the text of the top 100 websites obtained by using the search string "data analytics" using a search engine API, such as Google. One document is a webpage.

Note: this assignment will be much easier if you write a set of general functions that can take in any parsed input, instead of rewriting your code for each of i-vi.