Homework IV

You can submit in groups of 2!

All assignments need to be submitted via github classroom:
https://classroom.github.com/assignment-invitations/27800f3262a7657bdef7da3728d12011
The homework is due 04/17/17.

The repository will contain the following files:

- **hw4_starter_notebook.ipynb** : the notebook to be used for providing your final analysis

Use travis to run your notebook and add an "assert" statement inside the notebook that ensures the outcome is the actual error you report for each of the tasks.
(Say, if you claim your algorithm is 90% accurate, add an assert statement that score returned by your model is at least 90%).
Travis will simply run the notebook and check that it ran without errors.

We recommend that you fork the homework repository and run Travis on your own repository - this way you don't have to wait for other students submissions to finish on travis.

Tasks

The task is to do text classification on a dataset of complaints about traffic conditions to the city of Boston. You can find the data here: https://data.boston.gov/dataset/vision-zero-entry
There are two goals:
First, try to predict the type of complaint ("REQUESTTYPE") from the complaint text.
Second, try to come up with a better categorization of the data into semantic categories.

1. **Data Cleaning [10 points]**
    - Load the data, visualize the class distribution. Clean up the target labels. Some categories have been arbitrarily split and need to be consolidated. Also remove duplicate data points.
2. **Model 1 [10 points]**
    - Run a baseline multi-class classification model using a bag-of-word approach, report macro f1-score (should be above .5) and visualize the confusion matrix. Can you interpret the mistakes made by the model?
3. **Model 2 [30 points]**

    - Improve the model using more complex text features, including n-grams, character n-grams and possibly domain-specific features.
4. **Visualize Results [10 points]**
    - Visualize results of the tuned model (classification results, confusion matrix, important features, example mistakes).
5. **Clustering [10 points]**
    - Apply LDA, NMF and K-Means to the whole dataset. Can you find clusters or topics that match well with some of the ground truth labels? Use ARI to compare the methods and visualize topics and clusters.
6. **Model 3 [30 points]**
    - Improve the class definition for REQUESTTYPE by using the results of the clustering and results of the previous classification model. Re-assign labels using either the results of clustering or using keywords that you found during data exploration. The labels **must** be semantically meaningful.
      The data has a large “other” category. Apply the topic modeling and clustering techniques to this subset of the data to find possible splits of this class.
      Report accuracy using macro average f1 score (should be above .53)
7. **Extra Credit [Up to +20 points]**: Use a word embedding representation like word2vec for tasks 3 and or task 6.