CSci 343 Fundamentals of Data Science
Challenge 5

Due: November 1, 2016 *before* 11:55PM
Points Available: 250 XP

Objectives:

- Learn about basic Nearest Neighbor Approximation
- Learn about linguistic preferences across the US
- Have fun!

Assignment:

We've been looking a lot a maps lately. In particular, we've been reconstructing missing data in maps (also images and plots) using *nearest neighbor approximation*. This method is often used to approximate trends in regions as determined by surveys or polls. One very good example is a linguistic survey originally administered by North Carolina State University (http://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6/).

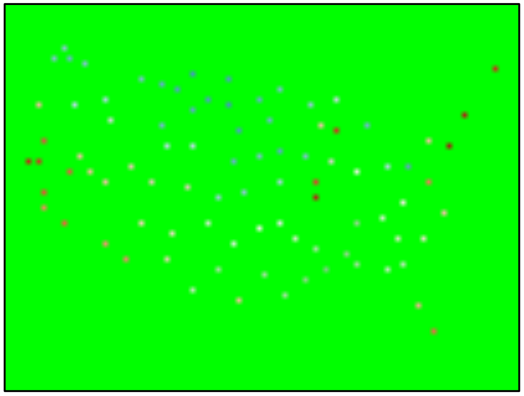
For this challenge, we will be using a subset of this data. Specifically, we're going to be investigating the results of the question "What is your generic term for a sweetened carbonated beverage?" Survey participants responded with one of four options: Soda, Pop, Coke, or Soft Drink. You have been provided (on the class data website) with two images. The first image is survey data with missing data colored green ([0.0, 255.0, 0.0]). In terms of our Lecture 21 slides, this will be your *Reference Image*. The second image is a US map template. The green areas represent where you need to fill in the image. This image, in terms of our Lecture 21 slides, will be your *Reconstruction Image*.

Your task is to implement the *Mean k-Nearest Neighbors Approximation* method. You are to use this method to generate a map of the US that shows the soda/pop/coke/soft drink word preferences across the country. In our class example, we used a copy of the Reference Image as the Reconstructed Image. For this assignment, you will use the blank US map without the reference samples included. Your program will need to take two command line parameters. The first parameter represents the number of neighbors you wish to sample (k-value). This parameter can be any integer greater than 0. The second parameter will be the name of the Reference Image file.

Reminder: Don't forget to use one of the more efficient methods of calculating distance! We talked about some of these in class. They are also discussed on the class wiki: http://www.cs.olemiss.edu/~jones/doku.php?id=csci343_nearest_neighbors

Here is an example of the input data and output image:

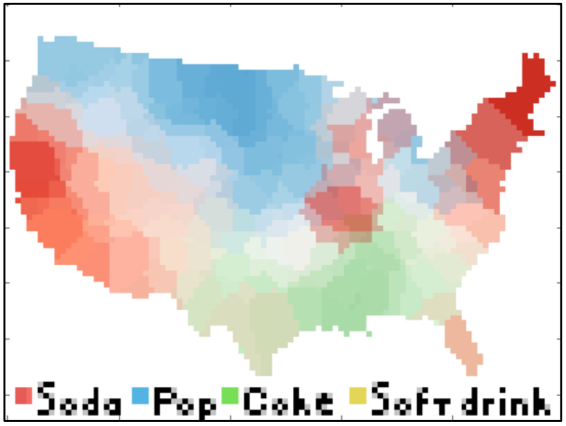
Reference Image



Blank Reconstruction Image

Soda
Pop
Coke
Soft drink

Fully Reconstructed Image:



Deliverables

Before you upload your code to Blackboard, you MUST demo your *working* project to the TA. Once you've done this, you can upload ***all your code*** and a saved image of your final output (named "reconstructed.png") to Blackboard as a single ZIP file. Name your ZIP file *spiritAnimal*.zip, where *spiritAnimal* is your class user ID (not your webID or ID number). Be sure to name your main source file "main.py". In a comment at the top of the file, include the following information.

- Spirit Animal User ID, Date the file was last edited, Challenge Number
- Cite any sources that you used as a reference for code, data, and content (including title and URL)