CPSC425: Assignment 1 (Python v2.7.15)

Alec Xu
38108130 (p7g9)

March 31, 2020

Question (4) Bag of Words Histograms

Using a vocabulary size of 100 (instead of the standard 50), I generated histograms to describe each category's bag-of-words representation. This is after normalizing to account for different lighting levels.

some histograms I found similar to each other were:

category 1	category 2
office	bedroom
living room	kitchen
industrial	inside city
bedroom	living roo

As we will see later in the confusion matrices, the squares at the cross-section of these categories are a lighter color than the other ones, barring the auto-correlations. Ideally only the auto-correlations would light up.

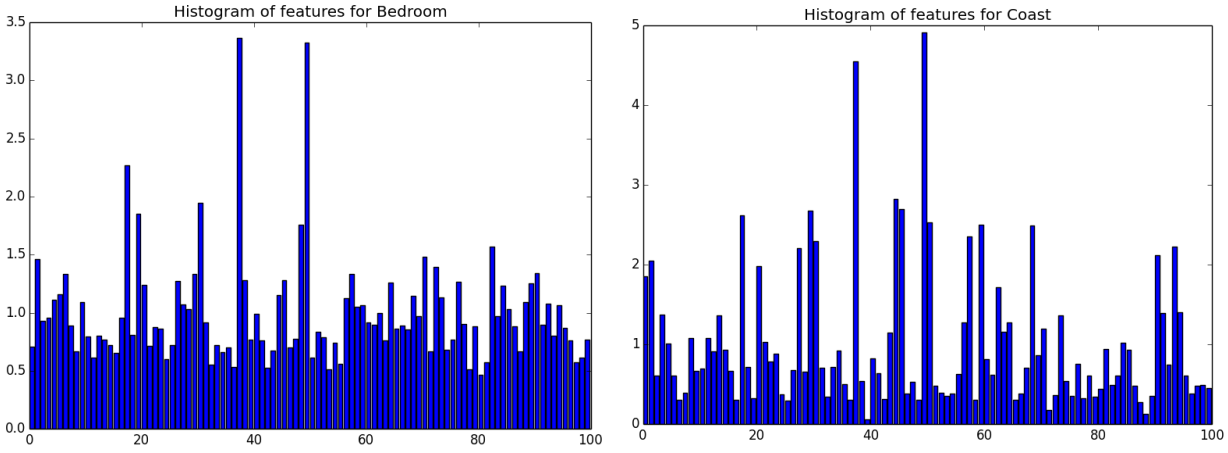

Figure 1: Bag of words represented as histograms with 100 "word" bins

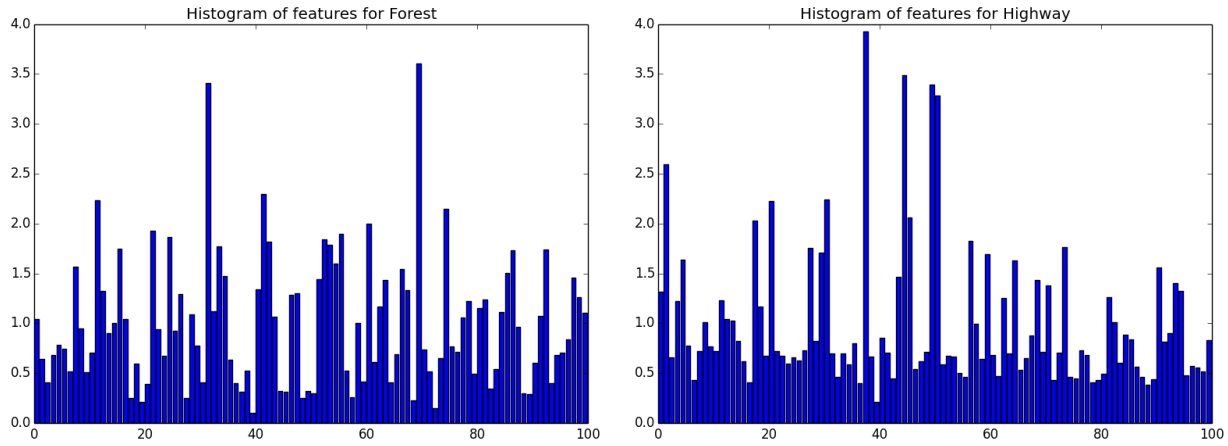

Figure 2: Bag of words represented as histograms with 100 "word" bins

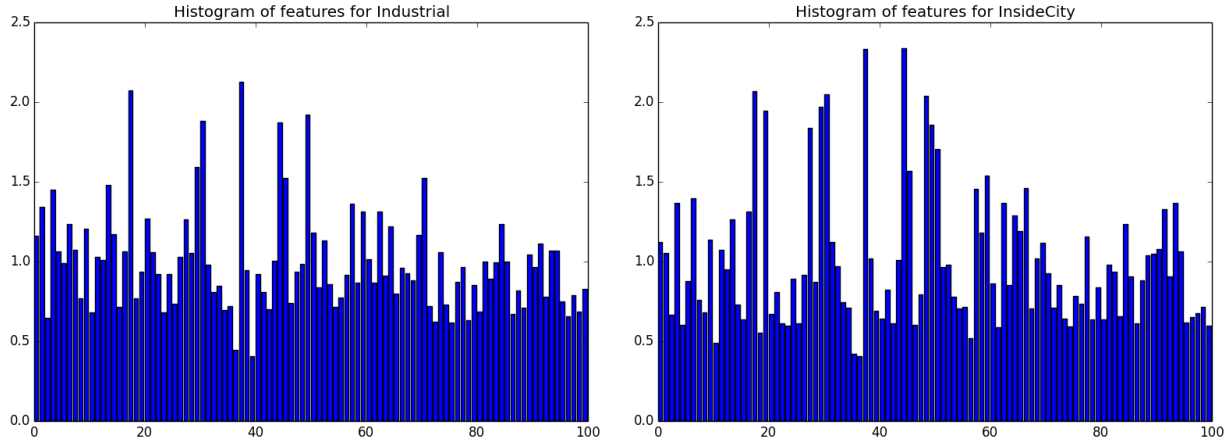

Figure 3: Bag of words represented as histograms with 100 "word" bins

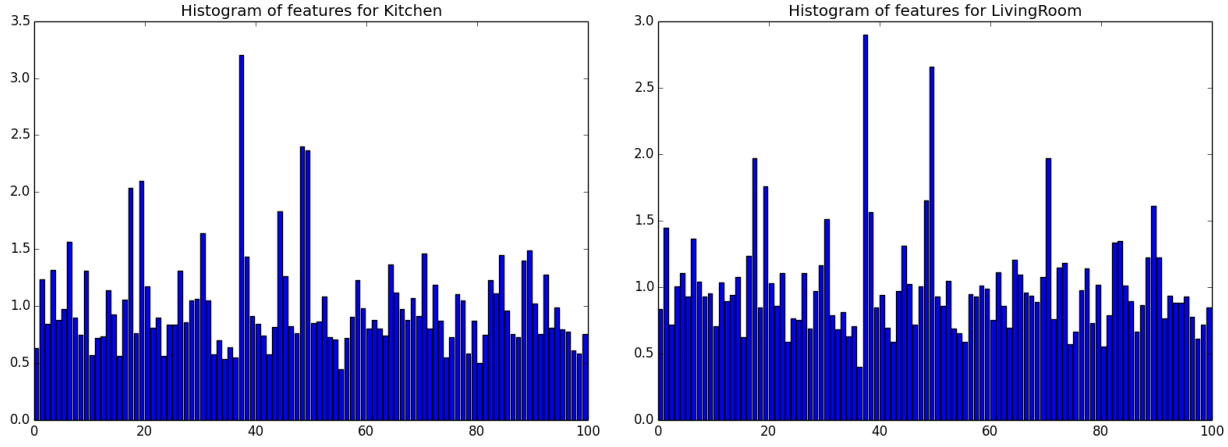

Figure 4: Bag of words represented as histograms with 100 "word" bins

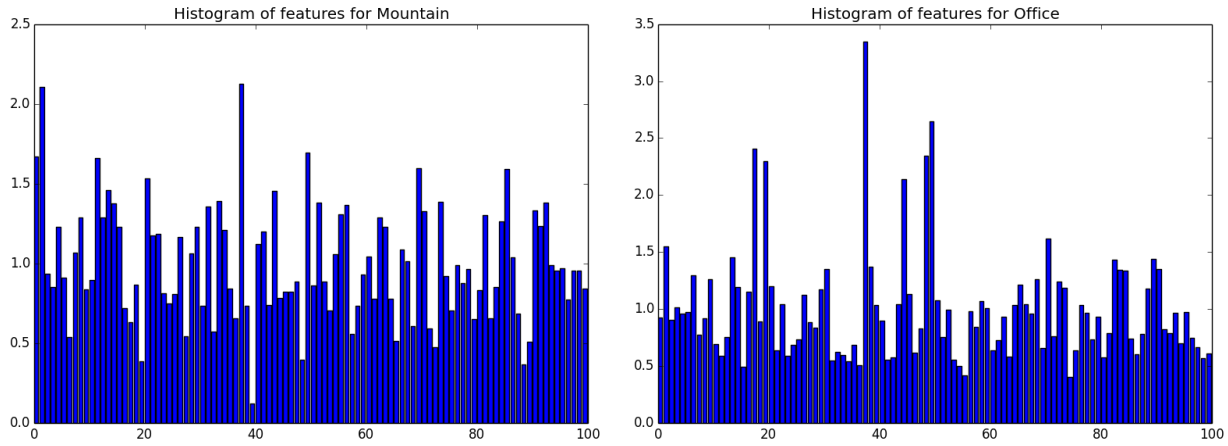

Figure 5: Bag of words represented as histograms with 100 "word" bins

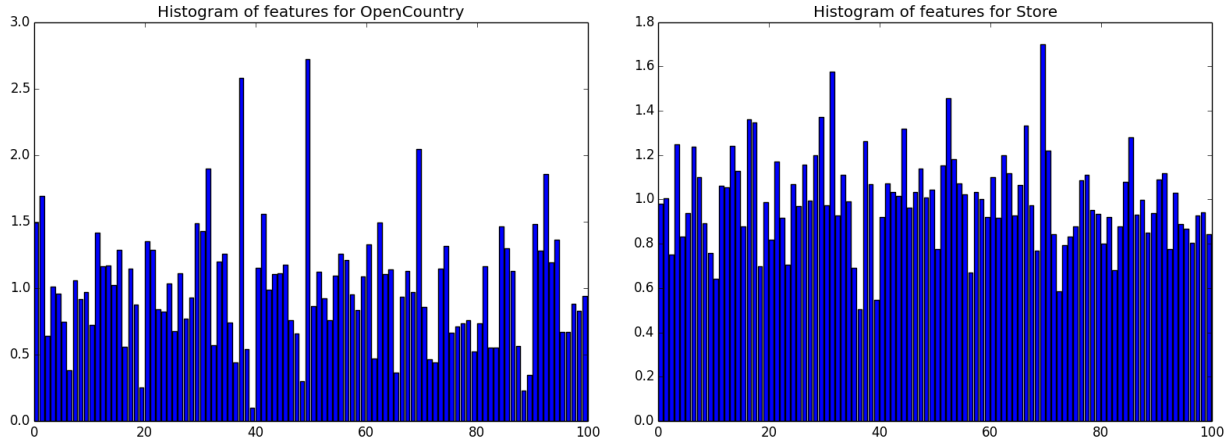

Figure 6: Bag of words represented as histograms with 100 "word" bins

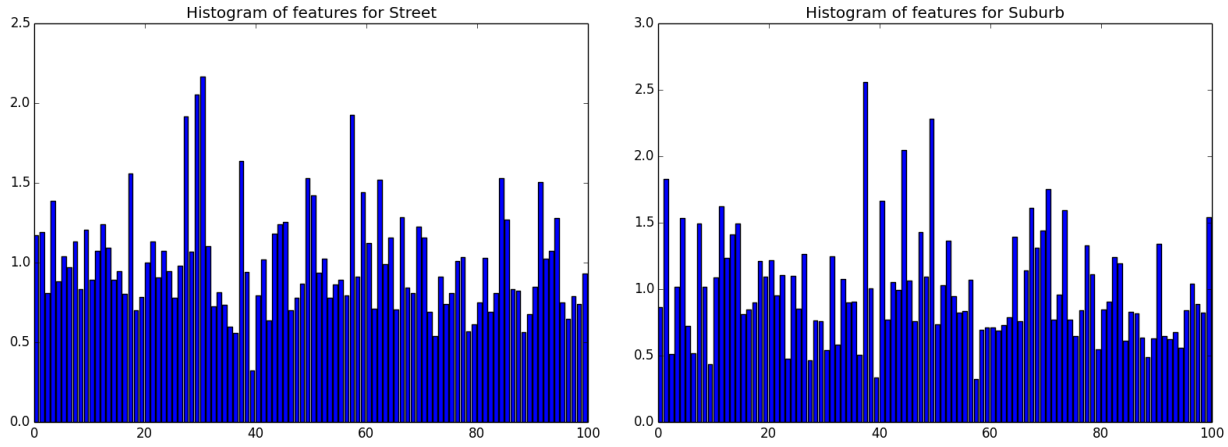

Figure 7: Bag of words represented as histograms with 100 "word" bins

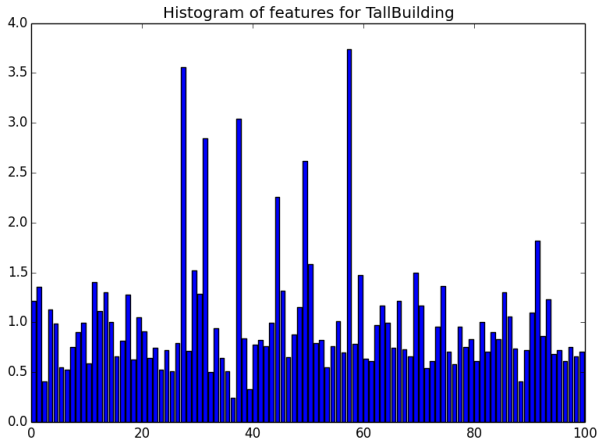

Figure 8: Bag of words represented as histograms with 100 "word" bins

Question (5) KNN Accuracy and Confusion Matrices

Vocab Size	# Neighbors	Accuracy (%)
100	1	38.44
100	10	37.33
100	100	32.44

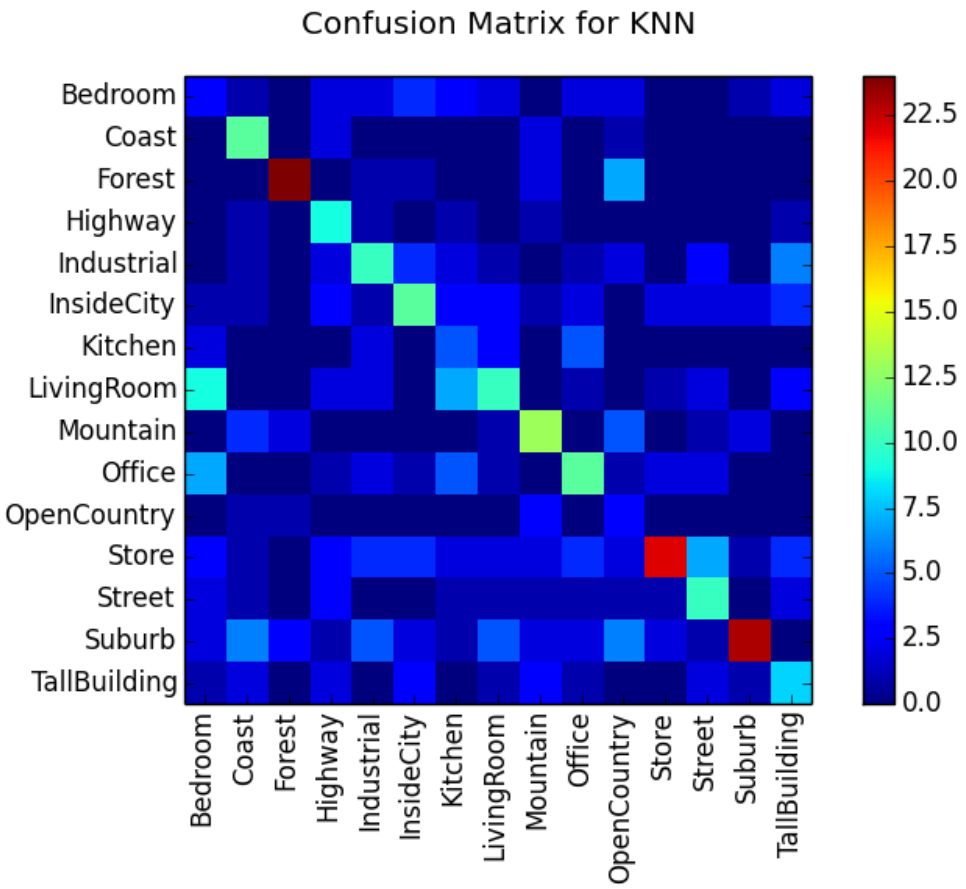

Figure 9: Confusion matrix of KNN with # neighbors=1, accuracy = 38.44%

I found that the less neighbors, the better results we achieved. This is not an expected result and could be due to the neighbors being widely spaced apart, and not sufficiently describing the problem domain. It could also Although our number of neighbors shouldnt be excessively high, such as 100, I expected the optimal number of neighbors to be more than 1.

Question (6) SVM Accuracy and Confusion Matrices

Vocab Size	Regularization (C)	Accuracy (%)
100	0.1	31.33
100	2	44.44
100	15	50.44
100	100	46.44

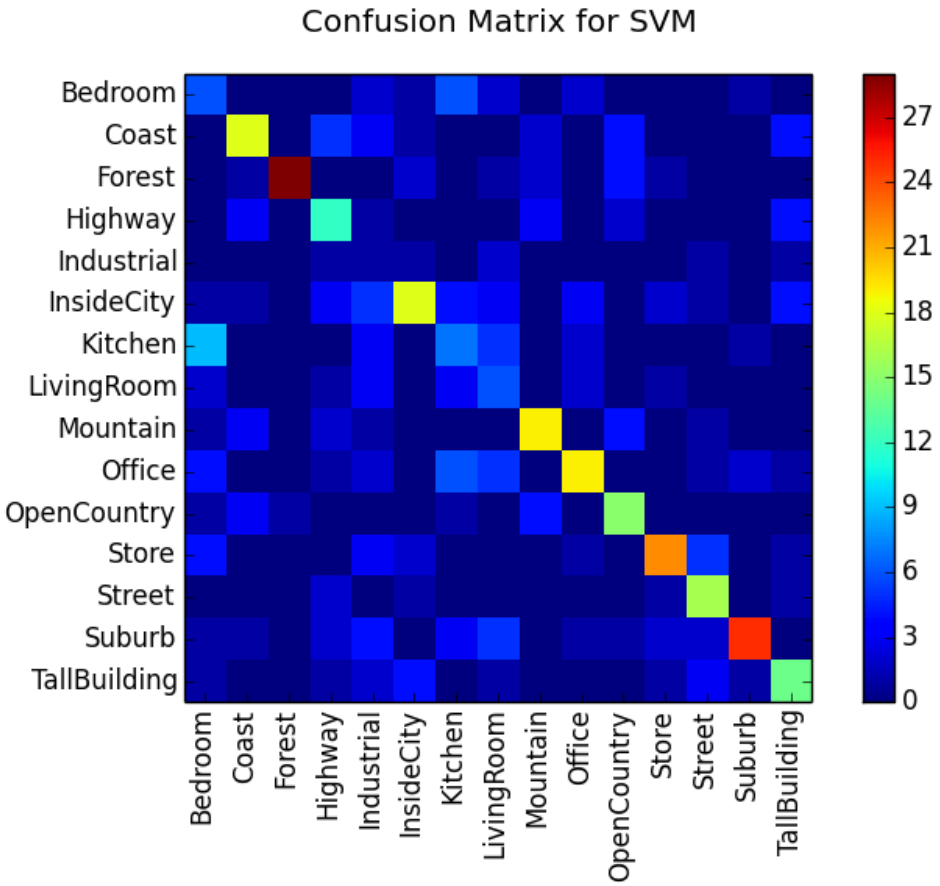

Figure 10: Confusion matrix of SVM with C=15, accuracy = 50.44%

I found that a modest regularization term for the SVM achieved the best result, at 50.44% accuracy. anything significantly smaller or larger gave a worse result. I also tried using different vocabulary sizes but 100 performed the best.