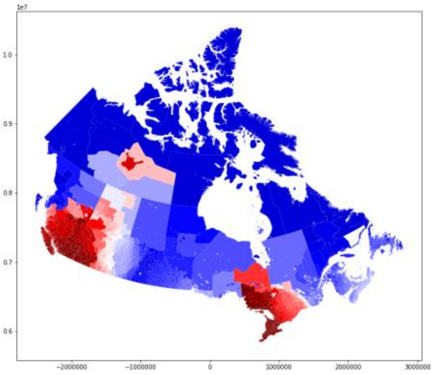
Raphael Ku
Uniqname: rapku
SI 671
Homework 3 Report

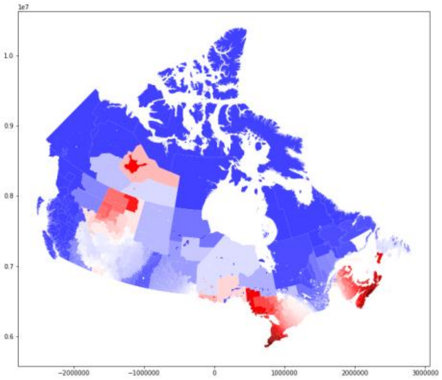
Part I. : Word maps

Evaluated using **z-sim** from pysal G_Local, and **geopandas**

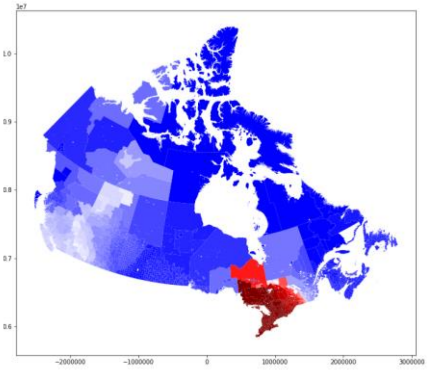
Word: snowboarding



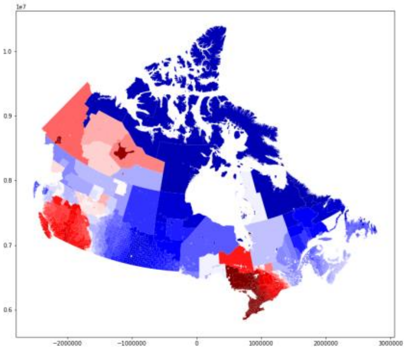
Word: sobeys (A grocery in Canada)



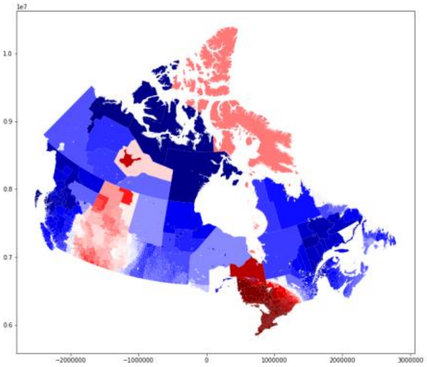
Word: timmies



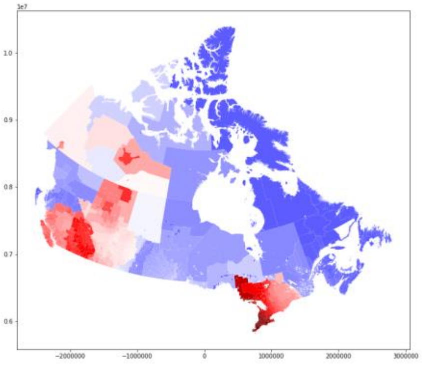
Word: motel



Word: tractor



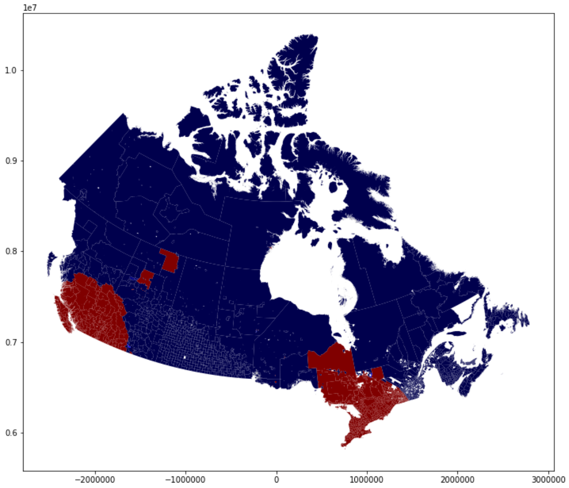
Word: Philippines
(note: my home country, know there should be some communities in Richmond area)



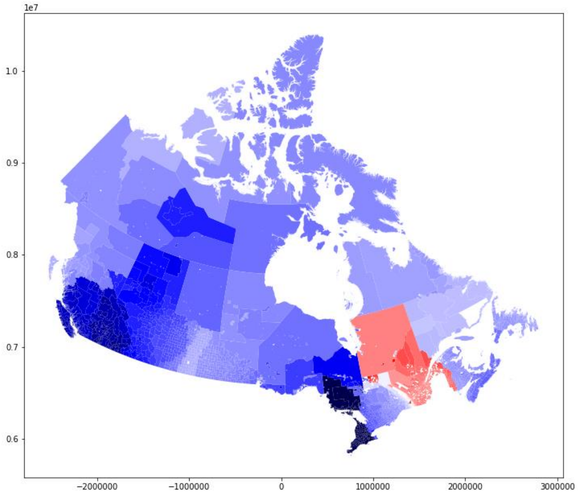
Part II: PCA Mapping (3 components)

PCA components to find set to 8 components, extracting 1st 3 principal components

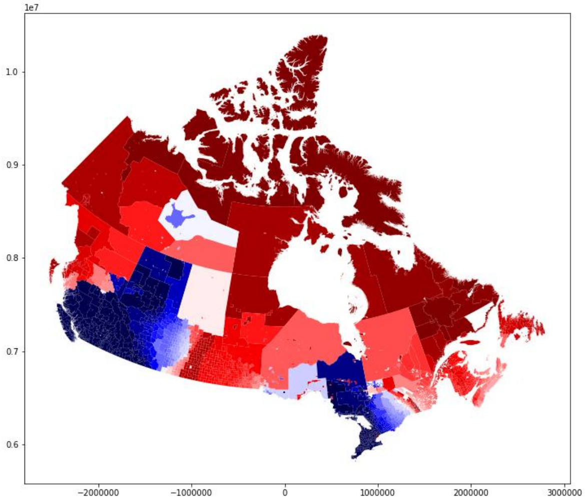
Component 1



Component 2



Component 3



Part III: PCA Analysis
Looking at the word weights provided for each principal component (using pca.components_ after running PCA), there are clear trends in the vocabulary used.

For **Component 1**, after sorting the weights, we see that the top (safe) words, like 'g2', 'friggen', 'highschool', 'gunna', 'boyfriends', 'probs', 'pita' are common in the principal component. This may indicate that the main users of Twitter in the indicated regions in the map are teenagers and young adults, given the casual/informal use of language in the top words.

Oppositely, words with the least weight are either professional words ('architect' - 0.007097), or highly specific names and brands.

Top 10 words for Principal Component 1 are as follows:

Word	**Weight**
cottage	0.022557
g2	0.022527
conservation	0.020662
von	0.020478
Hangout	0.019403
friggen	0.019402
chirp	0.018988
aloud	0.018592
highschool	0.018108
chirping	0.017860

Component 2's strongest words are clearly French, with the mapped projection of Z-scores highlighting the Quebec area reinforcing this. After reading through the meanings of some of the most weighted words for this component, it seems like the top words are referring to parking availability and locations in Montreal.

The first English word found in the sorted weight list is 'handler', with a 0.000549 weight, indicating the dominance of French vocabulary when users in the region are tweeting.

Top 10 words for Principal Component 2 are as follows:

Word	**Weight**
stationnement	0.985343
saint-denis	0.129744
faillon	0.077675
sainte-catherine	0.068582
disponible	0.026829
duluth	0.017670
saint-laurent	0.011777
boyer	0.008037
mont-royal	0.003312
vieux-port	0.003302

Lastly, the frequently used/heavily weighted words in **Component 3** seems to be a mixture of words from different industries (‘grass/rubbish’, ‘commercial/industrial’, ‘seizures/convulsions’), and location names. These regions may have been grouped together as a combination of rural areas, as well as industrial parks. This may indicate that the rural areas use Twitter as a means of emergency communication, and advertising for business-to-business operations.

In contrast, the least frequent/low weighted words in Component 3 seem to be focused on leisure and travel, with ‘resort, ‘campground’, ‘hiking’ in the lowest 10 words by weight. Interestingly, word variations of British Columbia have low weights in this component, reinforcing the map projection, which indicates the British Columbia area as a ‘cold zone’ for the vocabulary used in the indicated regions.

Top 10 words for Principal Component 3 are as follows:

Word	**Weight**
grass/rubbish	0.700135
sherbourne	0.345332
commercial/industrial	0.274632
trl	0.210386
hr332	0.189917
rescue-pumper-426	0.189776
shuter	0.170456
b/w	0.155165
pumper-324	0.130642
p332	0.128742