SI 671: Spatial Data Mining

Author: Koki Sasagawa
Date: 12-10-2018

Methods:

The main goal of this project is to analyze Canadian tweets over geographic data to detect any regional variations in word usage. The main libraries used for this task are the following:

- Geopandas - supports datatypes to handle geospatial data
- Pysal - spatial analysis functions (Local Getis-Ord G*)

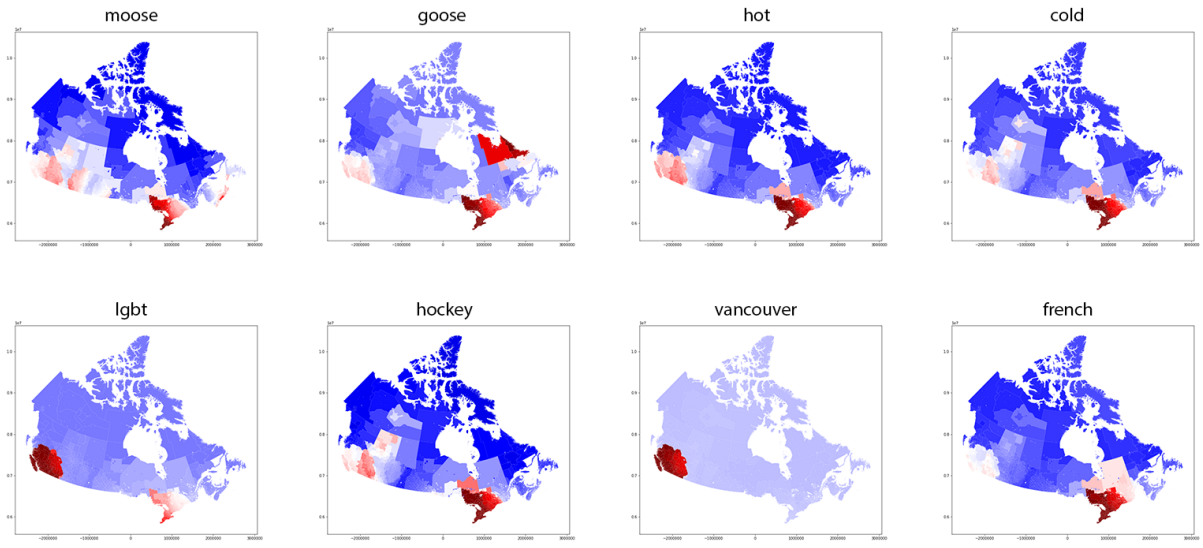
As Tobler's first law of geography states, "Everything is related to everything else, but near things are more related than distant things." This project explores spatial autocorrelation to understand how similar closer objects are to other nearby objects.

Compute Local Getis-Ord G* for select terms

The Local Getis-Ord G* algorithm allows us to analyze hot spots by calculating the Z-score which is used to determine where features with high or low values cluster spatially. The algorithm analyzes each feature within the context of neighboring features. Positive Z-scores indicate clustering of high values (hot spots colored red) whereas negative z-scores indicate clustering of low values (cold spots colored blue).

The Local Getis-Ord G* was calculated for the following terms to see if there was any interesting regional clustering (figure 1). The terms are moose, goose, hot, cold, LGBT, hockey, Vancouver, and French.

Figure 1



We can observe that the word ‘moose’ lights up several regions in the lower half of Canada. It is possible that the habitat range of moose is within these regions, which leads to an increase in sightings by people.

The word ‘goose’ shows slightly positive scores in the lower part of the west coast and high positive scores in the metropolitan region (Toronto and Montreal) and the upper north region. The high scores in the metropolitan area could be due to a concentration of stores that sell the popular brand Canada Goose as shown in figure 2.

Figure 2

Image source: Google maps

The word ‘hot’ has a high score in west coast Canada and metropolitan area including Toronto and Montreal. This could be due to close surrounding bodies of water which can cause summers to be humid in these areas. Similar to 'hot', the word 'cold' appears in many of the overlapping areas. However, the region on the west coast bordering the ocean has a negative score. This could be to due to the oceans which cause milder winters.

The term ‘Vancouver’ shows strong clustering in the west coast of Canada where the city is located. The word is not really used in the rest of the country.

‘French’ shows high clustering in regions of Canada with strong French influence such as regions with Montreal and Quebec.

The term ‘Hockey’ has a high score in regions where hockey teams are present, as shown in figure 3.

Figure 3

Image source: billsportsmaps.com

Lastly, the term ‘LGBT’ shows high scores in metropolitan areas such as Montreal, Vancouver, and Toronto. Cities attract various demographics and people from different backgrounds, which could lead to an increase in discussion/debates about social topics around sexuality and equality which have been drawing heavy attention in recent years.

It is important to mention that high scores for the selected words, with the exception of the ‘goose’, all appear in the lower half of Canada. Referring to the population distribution map of Canada, we see that population density is much heavier in the regions bordering the United States (figure 4). More people leads to more potential Twitter users, thus produce a higher number of tweets than less populated areas of Canada.

Figure 4

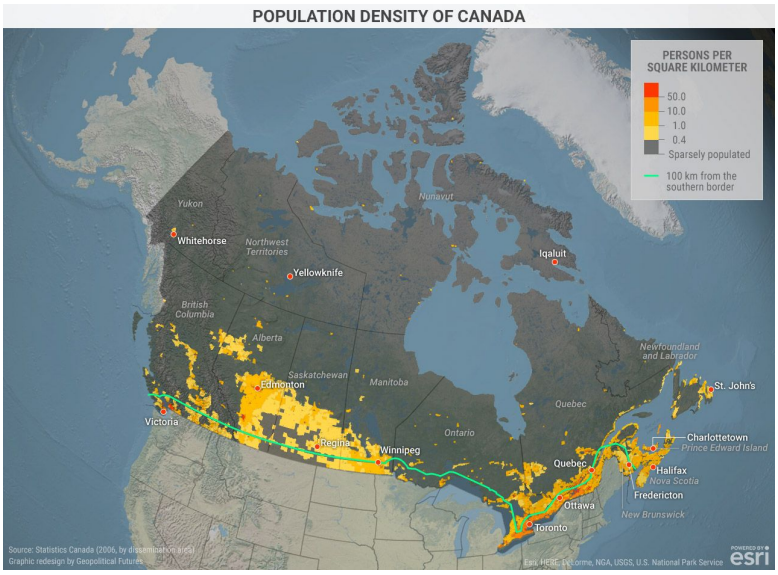
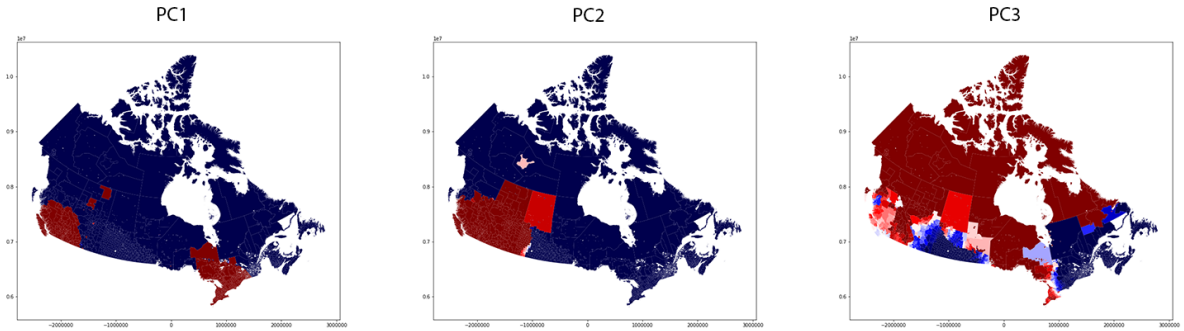


Image source: Geopolitical Futures

Compute Local Getis-Ord G* for all words then compute PCA

The Local Getis-Ord G* was computed for all words in the dataset, then Principal Component Analysis was computed to find the top 3 principal components that capture the most amount of variation in terms.



For each principal component, the top 15 features with the highest correlation were obtained. The following table summarizes these features:

PC1	PC2	PC3
cottage	british...	composite
g2	hiking	modis
conservation	resort	tile
von	kesler	gmt+0000
chirp	okanagan	processed
g1	hike	utc
chirping	campground	server
aloud	british	jaina
friggen	mountains	droid
defiantly	columbia	cst
gunna	bc	flood
highschool	fraser	han
tn	mountain	published
boyfriends	tug	cdt
timmies	b.c.	jedi

From PC1, I observed many instances of regional slang such as cottage (used to refer to camp), chirp/chirping (Canadian slang for talking smack), timmies (a fast-food coffee chain which got its name from a famous Canadian hockey player Tim Hortons) to mention a few. The regions which light up red are the west coast of Canada and area surrounding major cities in Toronto and Quebec which may attract a higher population of younger people.

PC2 shows an interesting trend of words associated with nature, such as resort, campground, hike, and mountains. The region that shows high score (red) is around the region of British Columbia, which is defined by its Pacific coastline and mountain ranges. Other words such as British, Columbia, bc and b.c. (short for British Columbia) are frequently present as users are more likely to mention their location in their tweets.

PC3 is very different from the previous 2 components, as it shows positive scores in much of the northern regions of Canada. There are mentions of what seem to be words associated with time. Gmt+0000 stands for Greenwich mean time, UTC which is short for Coordinated Universal Time, and CDT which stands for Central Daylight Time. There is also what looks like words associated with star wars characters such as Jaina, droid, Jedi, and Han. Due to lack of context, it is difficult to formulate any conclusions here, by my guess would be that these could be names used for ‘servers’ that the company ‘modis’ has in operation.