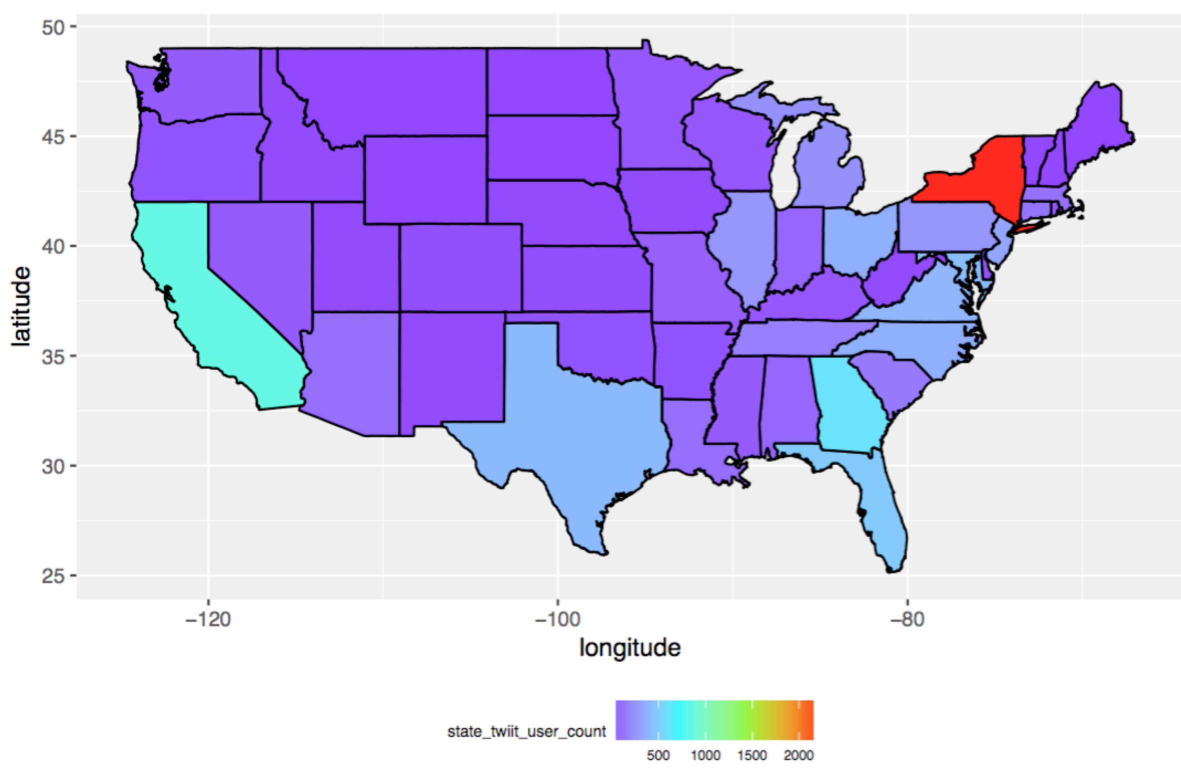
Twitter Data User geo-Location Prediction

Pei-Shien (Candace) Wu

North Carolina State University, Department of Statistics

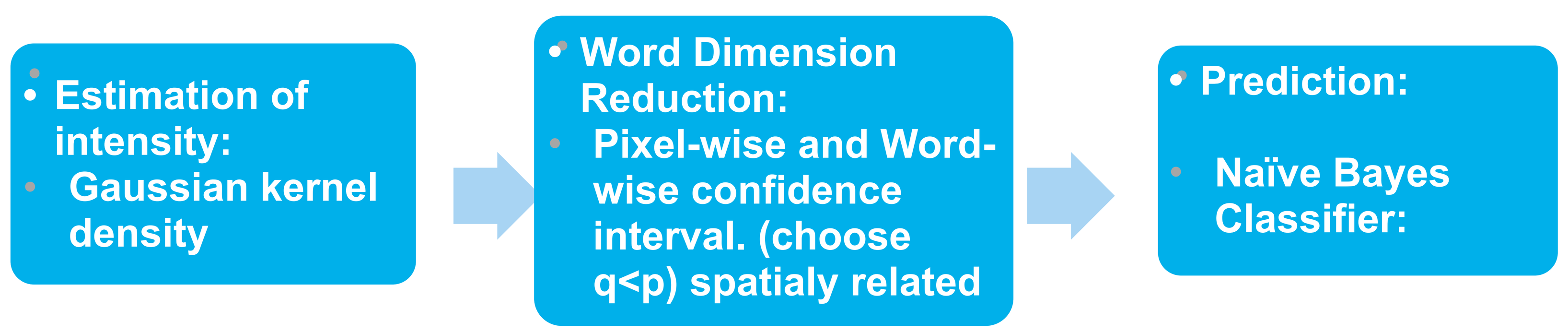
Introduction

- Text messages posted on microblog such as Twitter and Facebook has been known to be a good platform for detecting the outbreaks of events. The ability to predict the location of microblog users is useful in crisis management.
- Identifying Twitter user location based on the messages posted from the microblog users on Twitter is challenging due to the tremendous amount of unstructured textual data.
- The main goal of this project is to extract the unique language having spatial variability from p=5216 different characters, containing non-standard abbreviations, typographical errors, use of emoticons, irony, sarcasms and trending topics referred to as hashtags within the n = 8784 Twitter users in US.
- Furthermore, we predict the geo-location of tweets based on the characters having geographical variation.

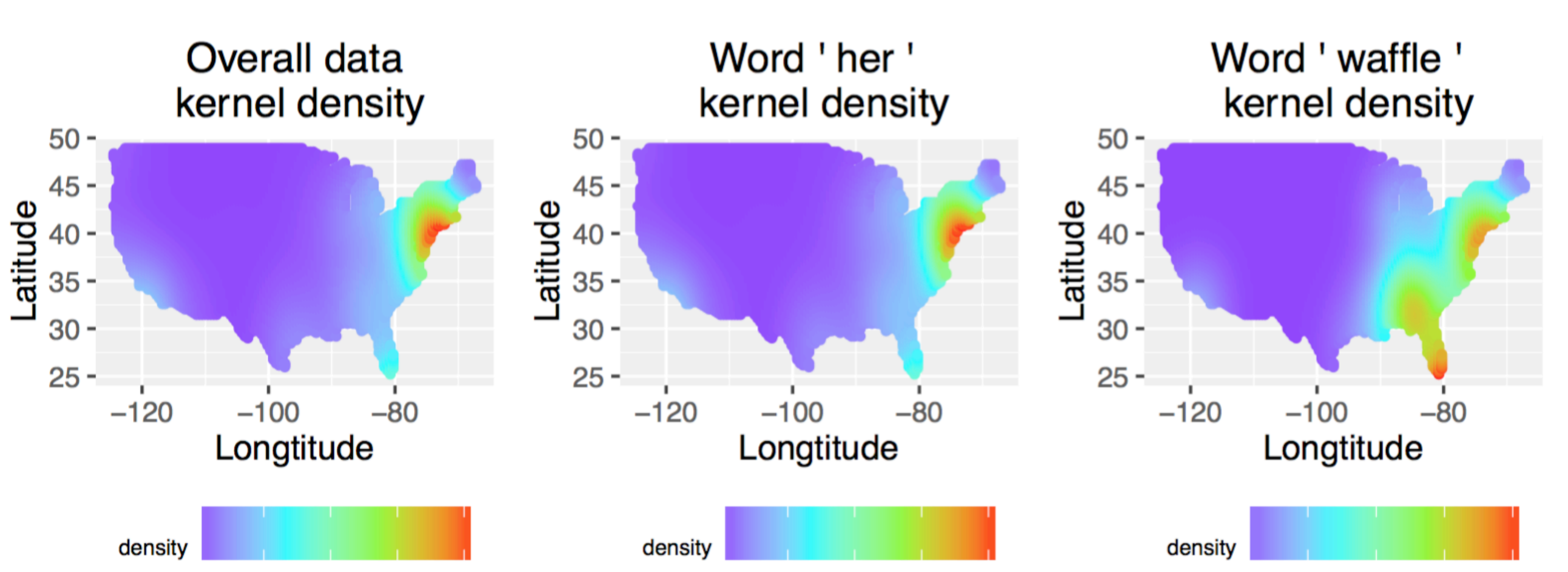


- Figure 1: Number of Twitter users in each state.

Method



- Estimate the intensity using normalized Gaussian kernel density function, and compare the intensity of each word with the overall data.



- Figure 2: Kernel Density Estimation.

Method

- **Pixel-wise confidence intervals: test the significance of each pixel.**

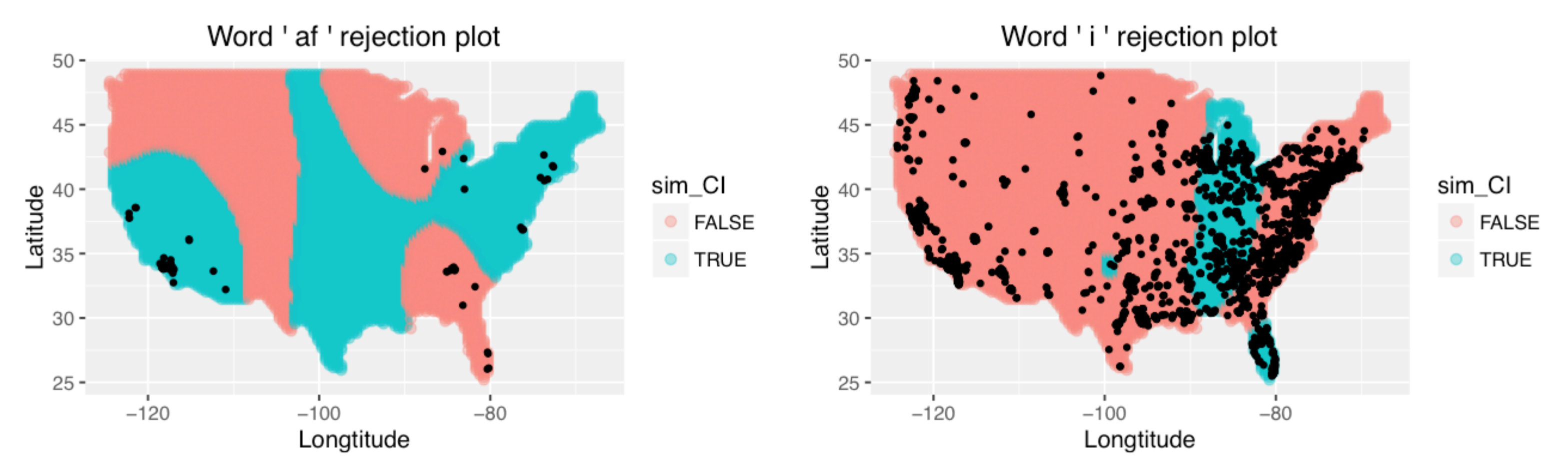
Let $\{x\}_{i=1}^{n}$ be the location of the overall data users, and $(\mathbf{P}_j(x) = \mathbf{Q}(x))$ are the kernel density of word j and overall data, for each word we test

$$H_0 : \mathbf{P}_j(x) = \mathbf{Q}(x) \; versus \; H_a : \mathbf{P}_j(x) \neq \mathbf{Q}(x)$$

1. There are n_j users tweets using word j, so we simulate $s = 200$ data and in each simulation we randomly select n_j data points from the overall data. The simulated data in each simulation is $z = \{z_i, \dots, z_{n_j}\}$ where $\{z_i\}_{i=1}^{n_j}$ is the location of word.
2. Estimate each simulated data kernel density $\mathbf{Q}^*(z_i) \in \mathbb{R}^{128\times128}; i = 1, \dots s$.
3. Construct confidence interval for each pixel:$\mathbf{P}(x) \in (\mathbf{Q}^*_{0.025}(z), \mathbf{Q}^*_{0.975}(z))$ where $\mathbf{Q}^*_{0.025}(z)$ is the 0.025 quantile of the estimate simulated data kernel density.
4. For each word j, the pixel-wise rejection criterion is

$$\mathbf{S}(x) = \mathbf{1}_{\{\mathbf{P}(x) \notin (\mathbf{Q}^*_{0.025}(z), \mathbf{Q}^*_{0.975}(z))\}}$$

Therefore, $\mathbf{S}(x)$ is a rejection map where each pixels in the map is a logical variable representing whether each pixels has a significant difference with the overall data at $\alpha = 0.05$.



- Figure 2: Rejection Map.

- **Word-wise confidence intervals: test the overall image significance.**

1. Based on the estimated simulation data kernel density $\mathbf{Q}^*(z_i) \in \mathbb{R}^{128\times128}; i = 1, \dots s$ leave out one simulation data and use the remaining $\mathbf{Q}^{*-i}(z) =$ to construct confidence interval $(\mathbf{Q}^{*}{}^{-i}_{0.025}(z), \mathbf{Q}^{*}{}^{-i}_{0.975}(z))$ where $\mathbf{Q}^{*}{}^{-i}_{0.025}(z)$ is the 0.025 quantile of the estimate simulated data kernel density.
2. For each left out simulation data, count the number of rejected pixel,

$$y_i = \sum_z \mathbf{1}_{\{\mathbf{Q}^{*i}(z) \notin (\mathbf{Q}^{*}{}^{-i}_{0.025}(z), \mathbf{Q}^{*}{}^{-i}_{0.975}(z))\}}$$

3. Use the empirical distribution of the number of rejected pixel $y_i; j = i, \dots, s$ to find the number of rejected pixel R which control the significant level at 0.05 which is $P(y > R) = 0.05 \Rightarrow R = \eta_y^{0.95}$.
4. The word-wise rejection criterion is

$$w = \mathbf{1}_{\{\sum_x \mathbf{S}(x) > R\}}$$

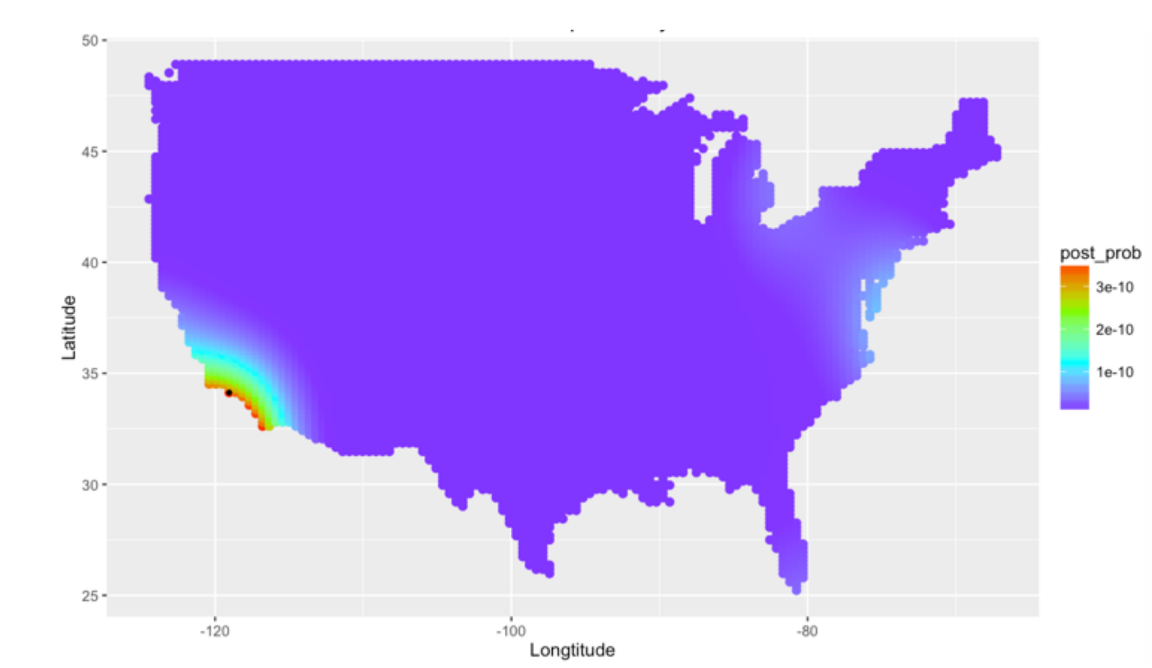
- **Naïve Bayes Classifier:**
  - Prior: The overall data normalized kernel density $\mathbf{Q}(x)$
  - Conditional probability:$\mathbf{P}_j(x) \forall j = 1 \dots, q$
  - Predicted tweets location using K phrases:$\hat{x} = \text{argmax}_x \mathbf{Q}(x) \prod_{k=1}^{K} \mathbf{P}_k(x)$

Results

- Applying the pixel-wise and word-wise test, we could extract q < p = 5216 words in our data, and the table below listed 10 words which have spatial variability, and another 10 words which have no spatial variability.

Word	Pixel-wise test (rejected pixels /total pixels)	Word-wise test	Word	Pixel-wise test (rejected pixels /total pixels)	Word-wise test
things	1146/9563	Fail to reject	mangoville	7849/9563	Reject
i	1100/9563	Fail to reject	bout	6937/9563	Reject
asap	893/9563	Fail to reject	yall	6006/9563	Reject
her	795/9563	Fail to reject	houston	5955/9563	Reject
different	786/9563	Fail to reject	nigga	4368/9563	Reject
saying	729/9563	Fail to reject	funn	4283/9563	Reject
my	398/9563	Fail to reject	lol	4272/9563	Reject
knock	300/9563	Fail to reject	waffle	3944/9563	Reject
waiting	186/9563	Fail to reject	freeway	3809/9563	Reject
now	154/9563	Fail to reject	faso	2285/9572	Reject

- Predict tweets location using Naïve Bayes Classifier. For example, the predicted location of tweets using word phrases "fasho freeway" is at (longitude, latitude)=(-119.049, 34.13)



- Figure 4: Posterior Probability Map.

Conclusion

- The simulation pixel-wise and word-wise confidence interval method take sample size of each words into consideration. However, there are p = 5216 words in our dataset, doing a simulation on a pixel-wise as well as a word-wise test might results in a time-consuming issue.
- Since we are using the kernel density of the overall data as prior and the north-east part of US has a relative high kernel density, might need to do scaling while predicting tweets location with only one word.

Reference

- Eisenstein, Jacob, et al. "A latent variable model for geographic lexical variation." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.
- Berman, Mark, and Peter Diggle. "Estimating weighted integrals of the second-order intensity of a spatial point process." Journal of the Royal Statistical Society. Series B (Methodological) (1989): 81-92.
- Waller, Lance A., and Carol A. Gotway. Applied spatial statistics for public health data. Vol. 368. John Wiley & Sons, 2004.