PS 1

Kanyao Han

Part I

1. Data access

1) Go to https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data, click the **Export** on the website, and choose a data format you want (csv, xml, JSON, RDF, etc.)
2) The data is stored on the Chicago Data Portal of the City of Chicago.
3) The curator is **the Chicago Department of Public Health**.

2. Citation

1. J. K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, J. Bhatt(2014). *Health Department Use of Social Media to Identify Foodborne Illness - Chicago, Illinois, 2013-2014*. Morbidity and Mortality Weekly Report, 63(32);681-685
2. D. E. Ho(2012). Fudging the Nudge: Information Disclosure and Restaurant Grading. The Yale Law Journal, 122(3):574:600

3. Data collection

This information is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010 to the present. The inspection are performed by staff from the Chicago Department of Public Health's Food Protection Program using a standardized procedure. The results of the inspection are inputted into a database, then reviewed and approved by a State of Illinois Licensed Environmental Health Practitioner (LEHP)

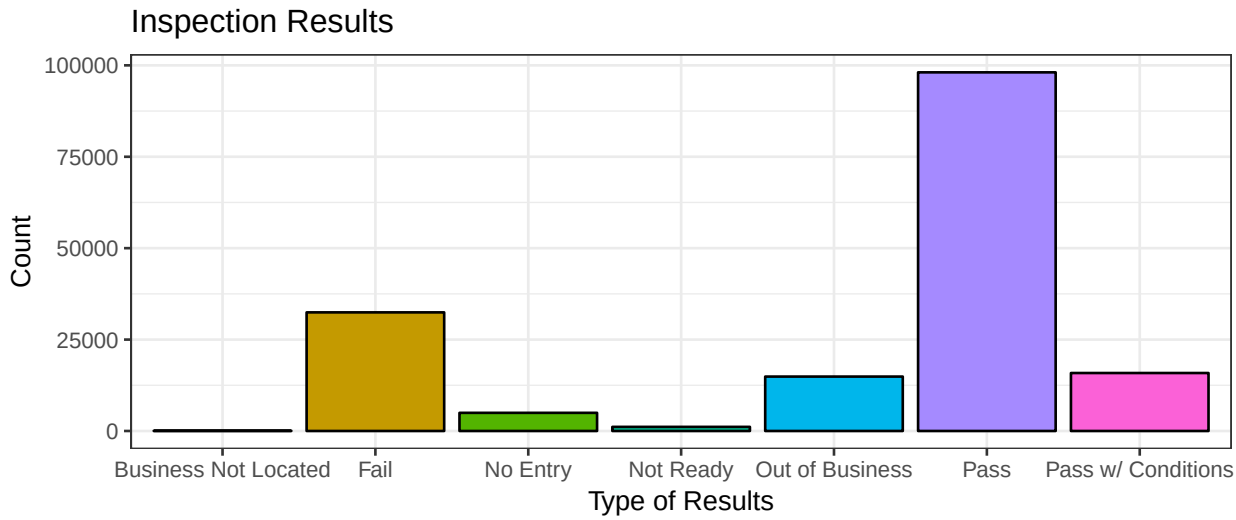
4. Descriptive statistics

Since the vast majority of the raw variables in the data is categorical, I transformed the format of some variables so that the descriptive stastics can provide more useful information. The final variables I select are zip code, Latitude, Longitude, Facility Type is restaurant or not, Risk 1 (High) or not, pass or not, pass with condition or not, fail or not, Inspection type is canvass or not, the number of violation is zero or not.

Table 1: Descriptive Statistics of some variables

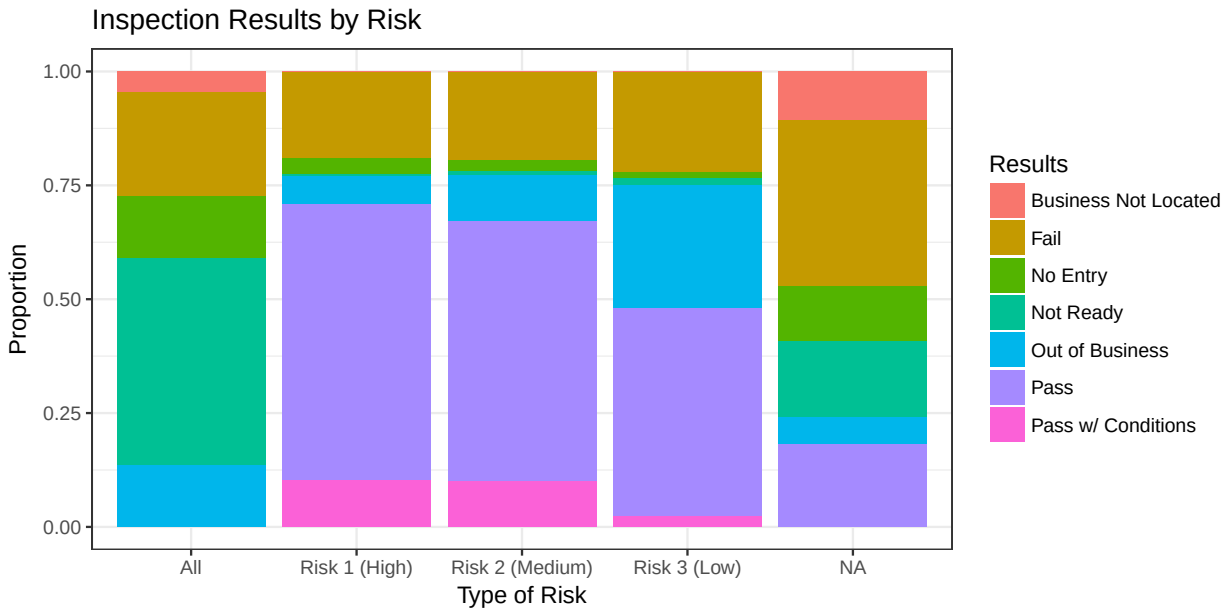
	mean	sd	median	min	max	range
Latitude	41.881	0.081	41.9	41.6	42.0	0.376
Longitude	-87.677	0.059	-87.7	-87.9	-87.5	0.389
Restaurant	0.681	0.466	1.0	0.0	1.0	1.000
Risk 1 (High)	0.703	0.457	1.0	0.0	1.0	1.000
Pass	0.586	0.493	1.0	0.0	1.0	1.000
Pass_with_conditions	0.095	0.293	0.0	0.0	1.0	1.000
Fail	0.194	0.395	0.0	0.0	1.0	1.000
Inspection_Canvass	0.531	0.499	1.0	0.0	1.0	1.000
zero_Violation	0.201	0.401	0.0	0.0	1.0	1.000

5. Visualization



The graph shows the vast majority of units (approx. 100,000) passed the inspection or got the pass with condition, but there were still a large number of units failed in the inspection.

6. Conditional description



If we look at the results by risk type, we can find from the graph that the units with low risk had lower pass rate and conditional pass rate, but a slightly higher fail rate. It is obvious that higher risk expected higher pass and conditional pass rates. Besides, the units with low risk has an incredibly high out of business rate. Finally, since the numbers of the units with all or NA in terms of risk are quite small, we should not compare them with the results in the groups of other three types of risk.

Part II

Grimmer, J. (2016). Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities

1) Research Question

The explicit research question is, by using a topic model, how legislators use communication to shape the type of representation (expressed priorities) they provide to constitutents, and how this definition of representtion changes in response to changes in institutional and electoral context (specifically, shifts in electoral pressure and changes in party control of Congress). The paper also has an implict question in methodology: how can we use computational tools to study representation.

2) Data

The author uses text as data method and a collection of every (nearly 170,000) House press release. They are from each House office from 2005 to 2010.

3) Theory

The paper doesn't have a formal theory since it uses an unsupervised learning model which means the author's goals are supposed to be hard to quantify. Because of this, the author take advantages of a topic model with a two-layer hierarchy which nests granular topics into a set of coarse topics and thus constructs the hierarchy of topics. All in all, this paper is much more data-driven than theory-driven. It is worth noting, however, that the author does have some assumptions that the number of coarse and granular topics are 8 and 44, and the granular topics can be nested in the coarse topics.

4) Classification

This paper is a combination of descriptive study and identification exercise. Unsupervised learning is typically used for exploration-oriented quantitative description since it does not have response. Therefore, for the topic modeling part, the author give us a description of the topics produced by the model (such as the changes of credit claming). Besides, the author also use identification exercise when he compute the correlation between supervised credit claim in a previous study and unsupervised credit claim in this paper, and regress the proportion of press releases in the farming category on the proportion of employed constituents who work in farming.

5) Computational methods

1. Method: This paper uses text as data method. The author uses a topic model with a two-layer hierarchy which nests granular topics into a set of coarse topics. Specifically, the author first transform the text content as numbers through a standard set of technique such as discarding word order, punctuation and capitalization, stemming the words, mapping words, removing the least and most frequently words and so forth. Then he constructs a hierarchy of topics through a topic model in which the granular topics are nested in the coarse topics, and describe as well as explain them. Finally, the author also validate the results in topic models by comparing it with the supervised learning results.
2. Results: House members expressed priorities lie on a credit claiming/position taking spectrum and it depends on who they present and responds to borad political changes. Specifically, after the 2008 election, Republicans abandon credit claiming articulate criticism towards Obama, while Democrats adopt credit claiming and defend the federal stimulus. It demonstrates that legislator will strategically

change how they communicate with constituents. However, it is worth noting that even after responding to the changing conditions, legislators largely maintain the same broad style. Besides, the author also provdes an methodological implication that the computational tools is useful for studying large number of text data of representation since the unsupervised learning results are highly correlated with supervised learning results in validation parts.

6) Suggestion

1. For topic modeling, the number of topic is very important and requires elaboration. However, the author does not give us the specific results of the initial experiments for results. I think he should display the results of the initial experiments. What's more, comparing different results from different topic-number selections in validation set might be more help to demonstrate the number of topic in this paper is reasonable.
2. Since the paper also focuses on methdology, I feel the author can talk much more about the advantages and limitations of topic modeling (unsupervised learning) by comparing it with supervised learning method in representation studies, rather than just concentrating on the positive side of the methods in the paper. It can make the readers futher understand how can we choose methods in specific situations.