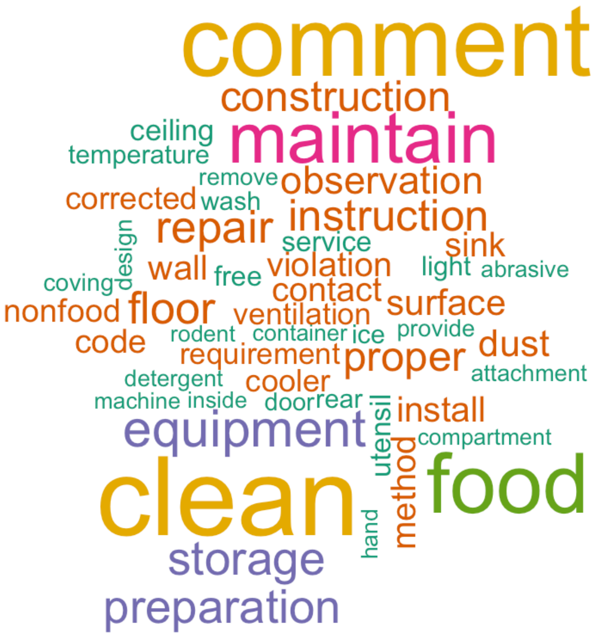
GROUP

Visulization-2-Word Cloud [Built in R]

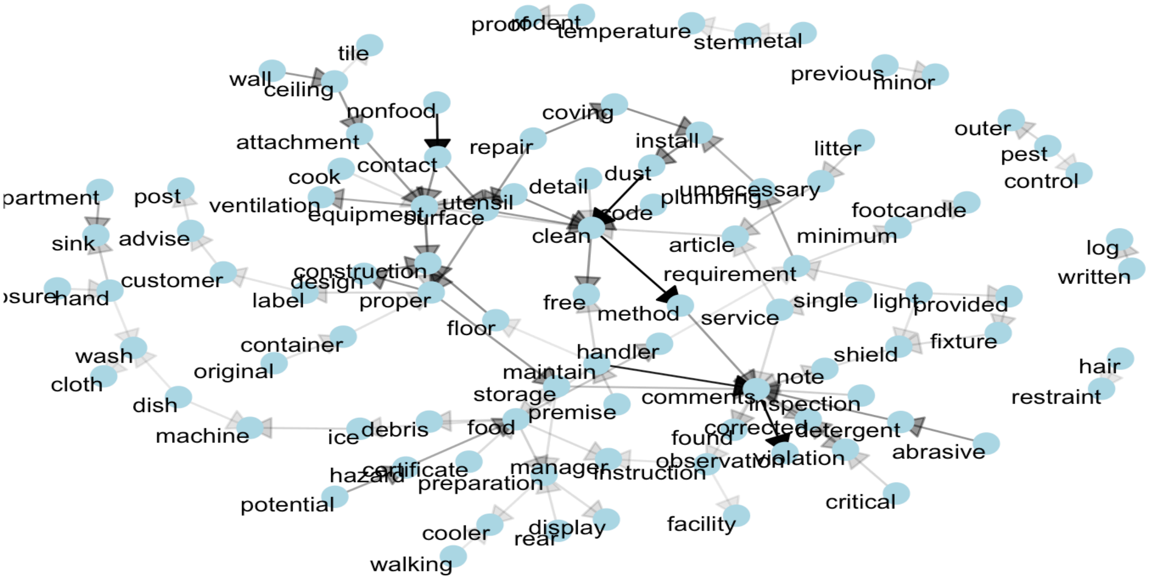
Word Cloud to visualized "Fifty" most important words in restaurant violations based on the "TF- IDF" scores. Higher the score, bigger the size of the text is.



- Word Cloud is an effective way to represent keywords in text data and the importance of each word is shown with font size and color.
- From font size and color, we found that "Clean" is obvious the most prominent term which would be the biggest problems in violations, in the meanwhile, words like "maintain", "equipment" and "preparation" in pink and purple are also big issues to which restaurants should pay attention.
- Look through violation content in dataset, we found that "food" mostly appear in names of certain rules about violations and "comments" would appear only when there are complementary things beyond rules. So these two big words here means that violation rules related to food and complementary notes are mentioned frequently.
- Smaller words in green and brown colors seem to show some details about big issues mentioned above. Enlightened by these words, we infer that there are maybe some stories to tell. For example, don't wash hands, don't clean kitchen properly can cause clean issue; ventilation, light and other machine problem can cause equipment issue. We will dig into these in in network graph later.
- This word cloud helps us get the quick idea that violations mainly about clean, food, equipment and preparation.

Visulization-3-Force Directed Graph/Network Graph [Built in R]

Force Directed Graph to visualize relationship between words which occur more than 1500 times in restaurant violations



- Force directed graph is a form of network emphasizing on complementary elements and which is good for digging in stories from these relationships.
- Each node represents a word in restaurant violations. Each edge reveals a connection between these two nodes.
- Thickness of edge represent the total number of times that two words appeared together and it shows how strong this connection is, the thicker the edge, the stronger the connection is. We can get some strong connection like “clean->method”, “dust->clean” (which have the most two thick edge) and know cleaning method and dust cleaning are problems that most restaurants have in inspections.
- We can see that there are five small centers (“equipment”, “clean”, “construction”, “comments”, “food”) that obviously have much more edges connected with other nodes, which means they are keywords and violations are mostly related to them. At the same time, these centers and nodes make up of five small groups which deliver more precise information from different perspectives.
- Directed edges helps us better understand these relationships/information: For example, for the group with “equipment” center, it may tell us that equipment related to cooking (like cooking utensil) and kitchen (like wall, ceiling) should be pay attention. For the group with “comment” center, we may get complementary information which is beyond regular violation rules, like service and observations about facilities that need to be corrected.
- This network graph tells detailed information about violations among words. In general, cleaning should be detail-oriented using correct methods; except cleaning, maintenance and service should be taken care according to the comments; food preparation and disposal should be proper dealt with; restaurant construction and equipment need to be repaired and maintained periodically.

INDIVIDUAL
Wenyi Yan

In the group, I take charge of data analysis of restaurant part, providing network graph making which is about restaurant violation keywords. In addition, I participated group discussion actively and came up with new ideas like to research on bad performing restaurants (with high risk level and high failure rate) in Chicago which has very practical meaning. In exploratory analysis part, I did some complementary graphs like interactive time series line graph of food inspection frequency in recent 3 years and heat map about distribution of risk level of establishments in recent 3 years to help group better explore the data, although they didn't show up in main body because of the limitation on graph amount.

Except network graph showed above, I also created many other graphs from different perspectives to dig more into restaurant data, such as bar graph which is to find Top 20 worst restaurants; choropleth which is to get the distribution of these restaurants; word cloud which is to refine keywords of violations; and leaflet map of restaurants broken down by fail and pass results which is interactive. I'll put the codes and whole graphs I did in Appendix B.

In this course, I learnt many data visualization techniques to discover insight in data and learnt how to use these techniques properly for different purposes. For example, I used word map to discover violation keywords and network graph to discover relationships between them. Besides, I learnt how to collect different visualizations together and tell a complete story about my findings. For example, I wanted to tell a story about bad performing restaurants in Chicago, so I used bar graph to find Top 20 bad performing restaurants, then used choropleth to discover the distributions of these restaurants with high risk level and high failure rate, and used word cloud and network graph on violation, exploring and analyzing what made them perform bad.

What's more, I learned a new domain of knowledge – NLP (Natural Language Processing) and a new statistics theory – TF-IDF (Term Frequency–Inverse Document Frequency) while doing word cloud and network graph. To process text data in "Violations" in the dataset, I used tm packages in R to remove stop words, lowercase and stem words, replace and delete inappropriate words. To refine the most important words in violations, I used TF-IDF to compute weights according to the frequency of usage inside an individual document as opposed to the entire dataset.