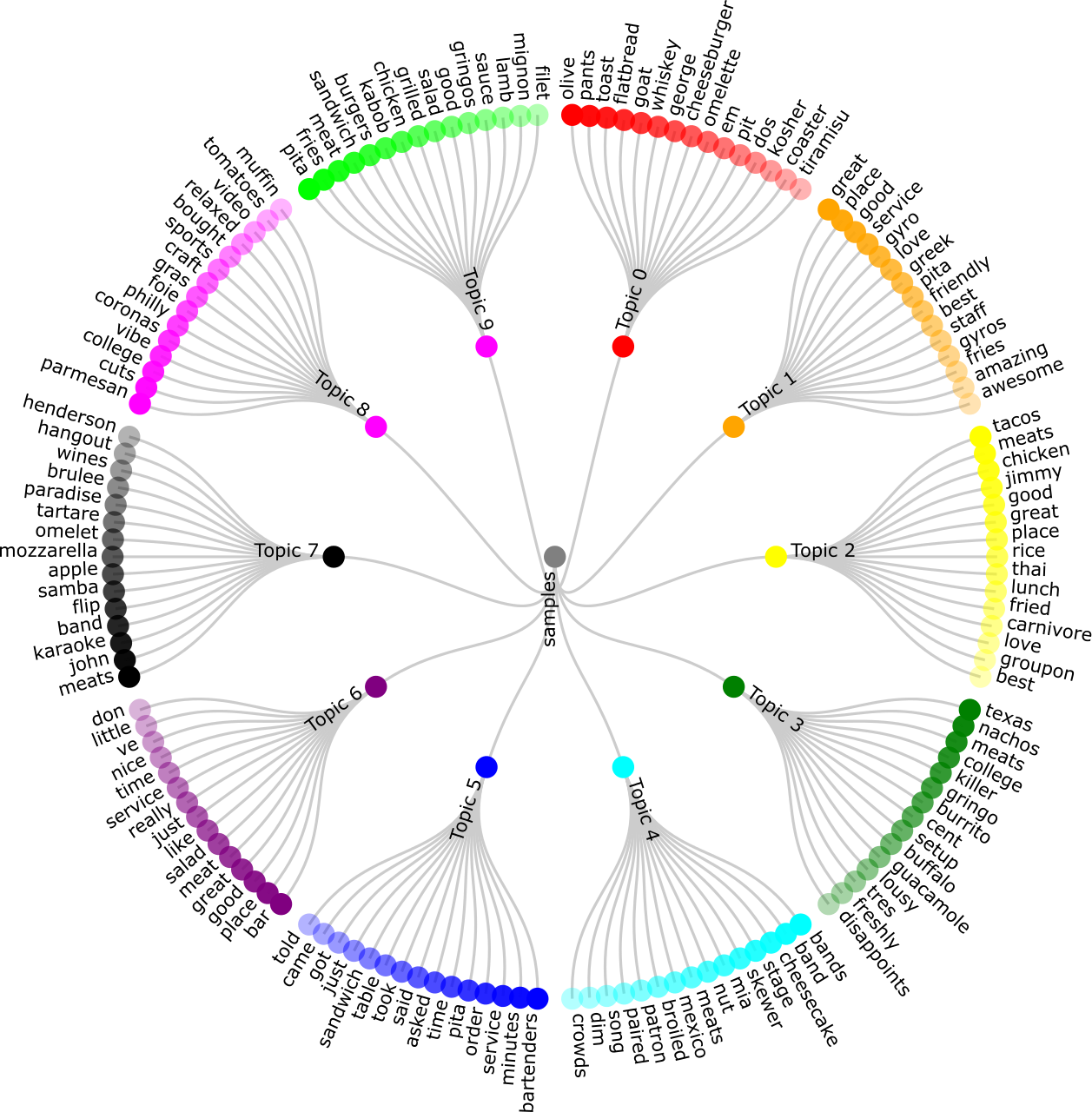
Task 1: Exploration of Data Set

# Task 1.1



## What

Topic model: LDA

Parameters: 10 topics

Dataset: 100000 samples reviews from the Yelp review

Visualization Tools: D3, InkScape

## Learned

If we check the combined frequent noun words based on the topics, we get:

1. Meats (4 times): people go to restaurant to eat meats, not veggies. A good restaurant should provide decent meaty meat, and people will talk about it.
2. Service (3 times): people are looking for experience different than home cooking. A good restaurant should not brand it as “homestyle” or “home cooking”
3. Place (3 times): a good and easy to find and access location is important to restaurant
4. Time (2 times): No one wants to starving for too long

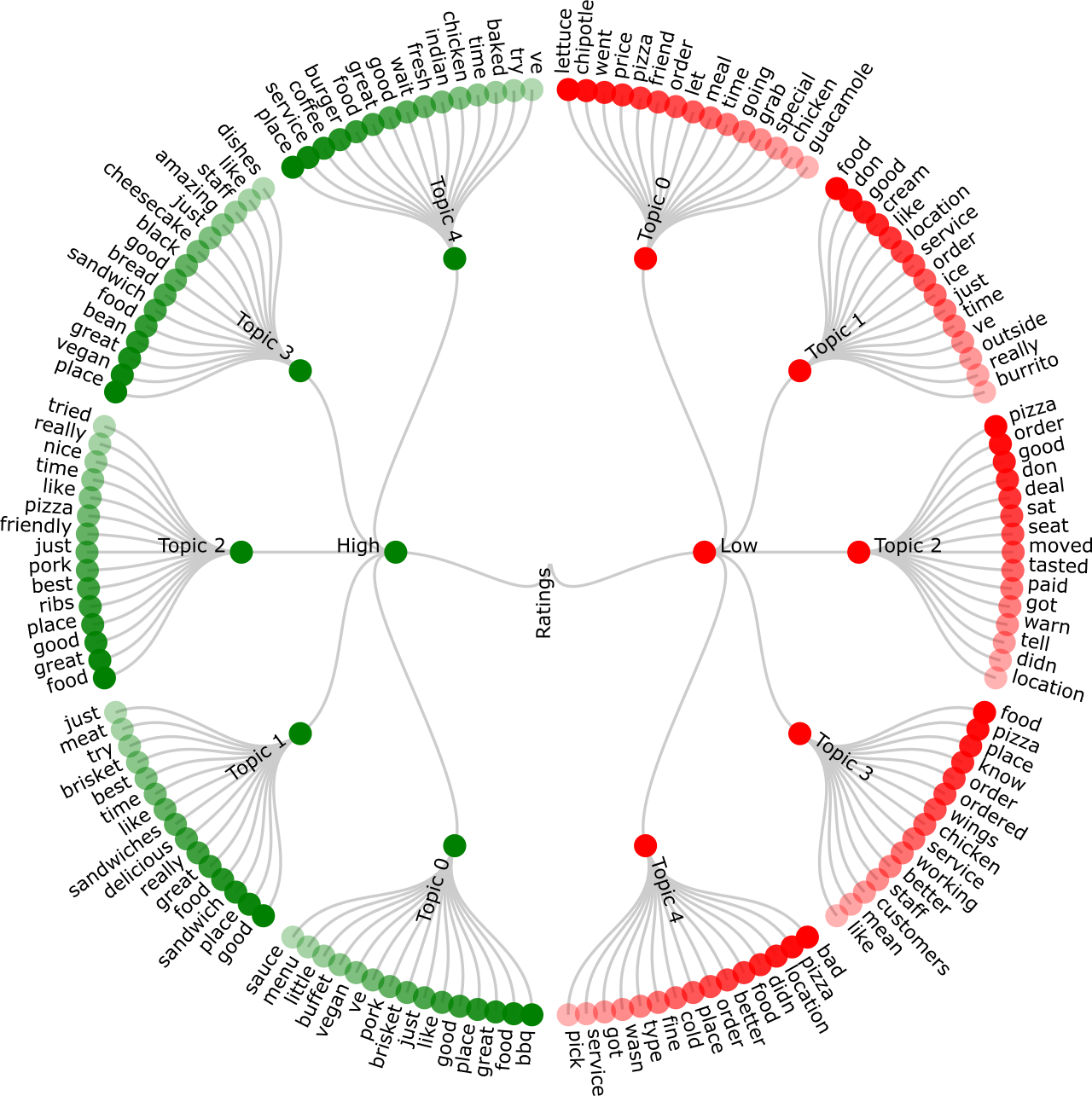
## How

1. Execute “python py27\_processYelpRestaurants.py --all” to generate sample data set “review\_sample\_100000.txt”
2. Execute “python py27\_ldaTopicModeling.py –f review\_sample\_100000.txt –o task1.1\_lda.txt –K 10” to generate topics
3. Convert topic data file task1.1\_lda.txt to Json format task1.1\_data.json
4. Edit topic data Json file task1.1\_data.json to include “color” and “opacity” attributes and corresponding values for each node
5. Edit the d3 js file task1.1.html to read and display it

## Where

<https://github.com/pauldeng/MOOC/tree/master/Data%20Mining%20Capstone/Task%201/Task%201.1>

# Task 1.2



## What

Visualization of restaurant reviews with high (5) and low (1) ratings.

Topic model: LDA

Parameters: 5 topics per each rating group

Dataset: Maximum 10000 samples review per each rating group from the Yelp review

High Ratings == 5

Low Ratings == 1

Visualization Tools: D3, InkScape

## Learned

With the knowledge of ratings, it is easier to guess the meaning of topics people talk about. If we check the combined frequent noun words based on the topics:

What we learned from high rating restaurants:

1. Food (5 times): obviously, good restaurants provide decent food.
2. Place (5 times): probably location is also equally important to good restaurants.
3. Time (3 times): do not let people wait too long.

What we learned from low rating restaurants:

1. Pizza (4 times): it seems to make a good pizza is not that easy
2. Food (3 times): needs to be improved
3. Location (3 times): maybe below average restaurants is not easy to access
4. Service (3 times): service also needs to be improved

## How

1. Extract both good and bad rated restaurant reviews from “yelp\_academic\_dataset\_review.json” based on “restaurantIds2ratings.txt”
2. Execute “python py27\_ldaTopicModeling.py –f badReviews.txt –o task1.2\_Bad.txt –K 5” to generate bad topics
3. Execute “python py27\_ldaTopicModeling.py –f goodReviews.txt –o task1.2\_Good.txt –K 5” to generate good topics
4. Convert topic data files to Json format
5. Edit the d3 js file task1.2.html to read and display it

## Where

<https://github.com/pauldeng/MOOC/tree/master/Data%20Mining%20Capstone/Task%201/Task%201.2>