Task 1: Exploration of Data Set

I've completed the first task using python language and jupyter notebook tool. In the performed work I used the following five data sets.

1. all yelp reviews,
2. reviews with one star,
3. five stars reviews,
4. restaurant reviews,
5. shopping reviews.

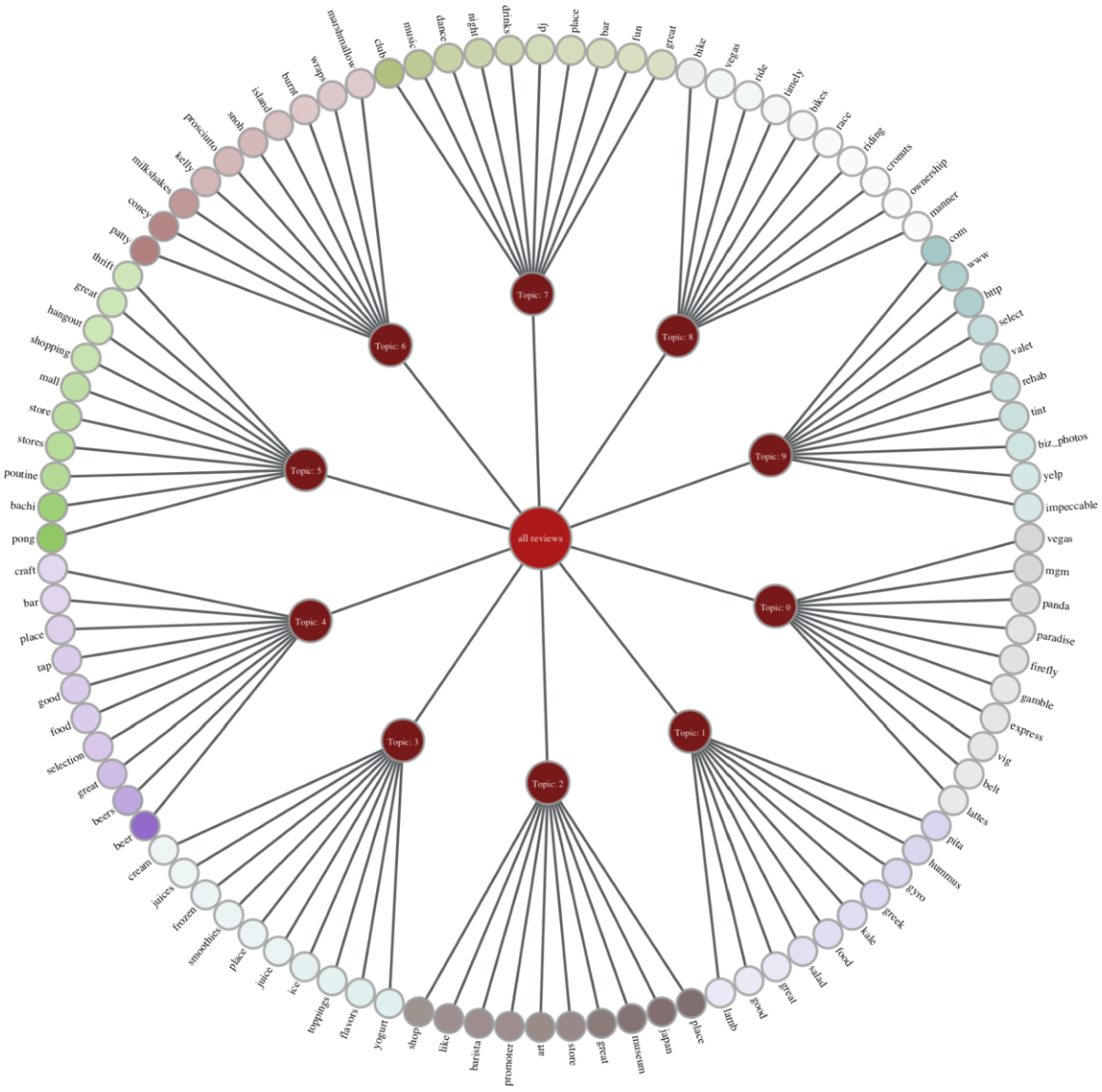
For data extraction and preparing I used Apache Spark SQL (https://spark.apache.org/ docs/latest/sql-programming-guide.htm l).

With help of this framework I managed to load lazy all reviews without experiencing issues with limit of memory. Also it help to extract one/five star reviews and fetch reviews related to restaurants business and to shopping business.

For topic analysis I used **gensim** and **sklearn (TF-IDF)** python packages. In gensim I chose a multi-threading implementation of LDA model construction. For visualization I chose two representation tree of topic-words and bar diagram (for all reviews dataset). Tree representation was based on python **graph_tool** library (https://graph-tool.skewed.de/) and bar chart was drew with help of **pyplot** package (http://matplotlib.org/index.html)

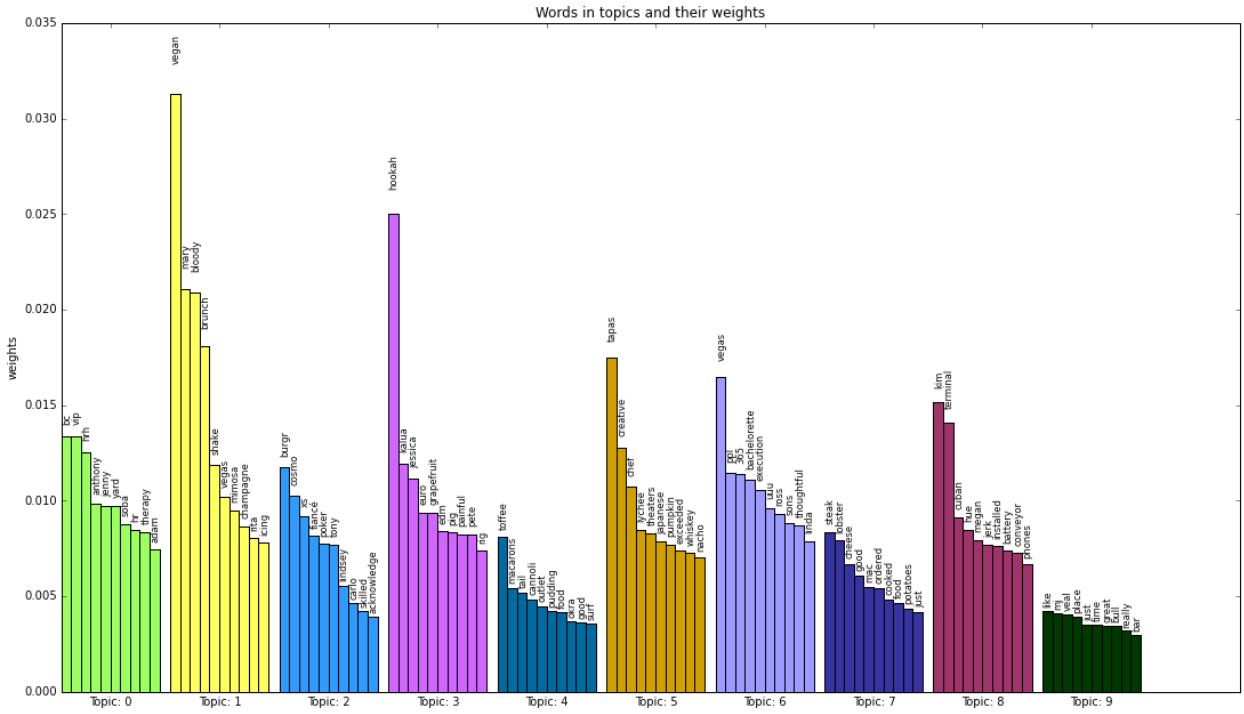
Task 1.1

LDA-model all reviews

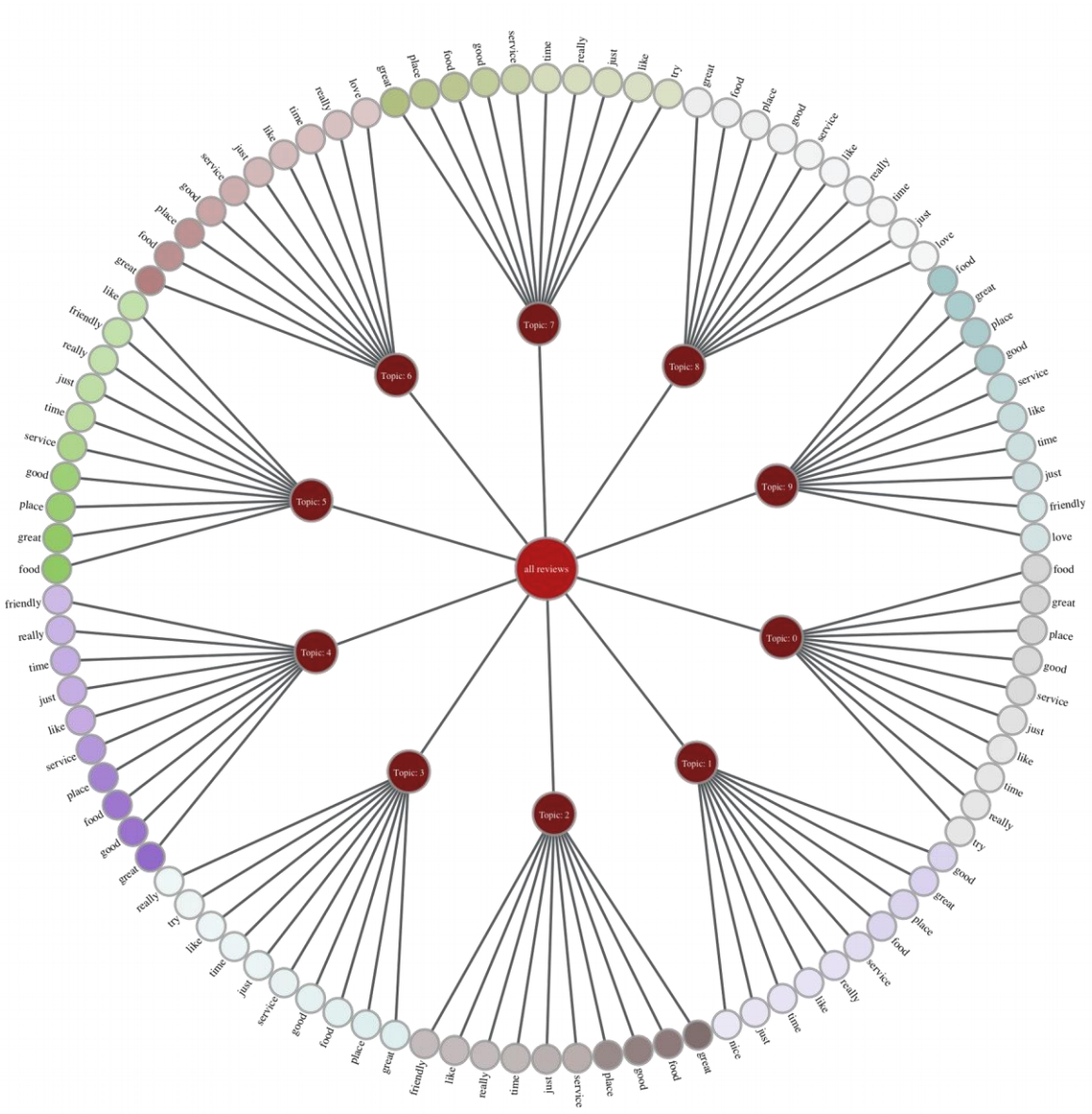


The visualization of the above gives an idea of the first 10 Topics. The intensity of color depends on weight of each word in the topic. Based on words some topics can be clearly assigned to a specific area, for example Topic #7 is obviously about night clubs: 'dj', 'drinks', 'nigth', 'dance', 'club', 'fun' and 'bar'. Topic #4 contains words from bars specific lexic: 'bar', 'tap', 'craft', 'beer', 'beers'. Topic #3 is more about beverages or ice while Topic #1 can be assigned to dishes topic.

Another visualization is built below. For this we use bar chart. Each topics presented as a sequence of bars and there is a gap between two topics. The visualization displays how weights are expanded across first 10 words in each topic.

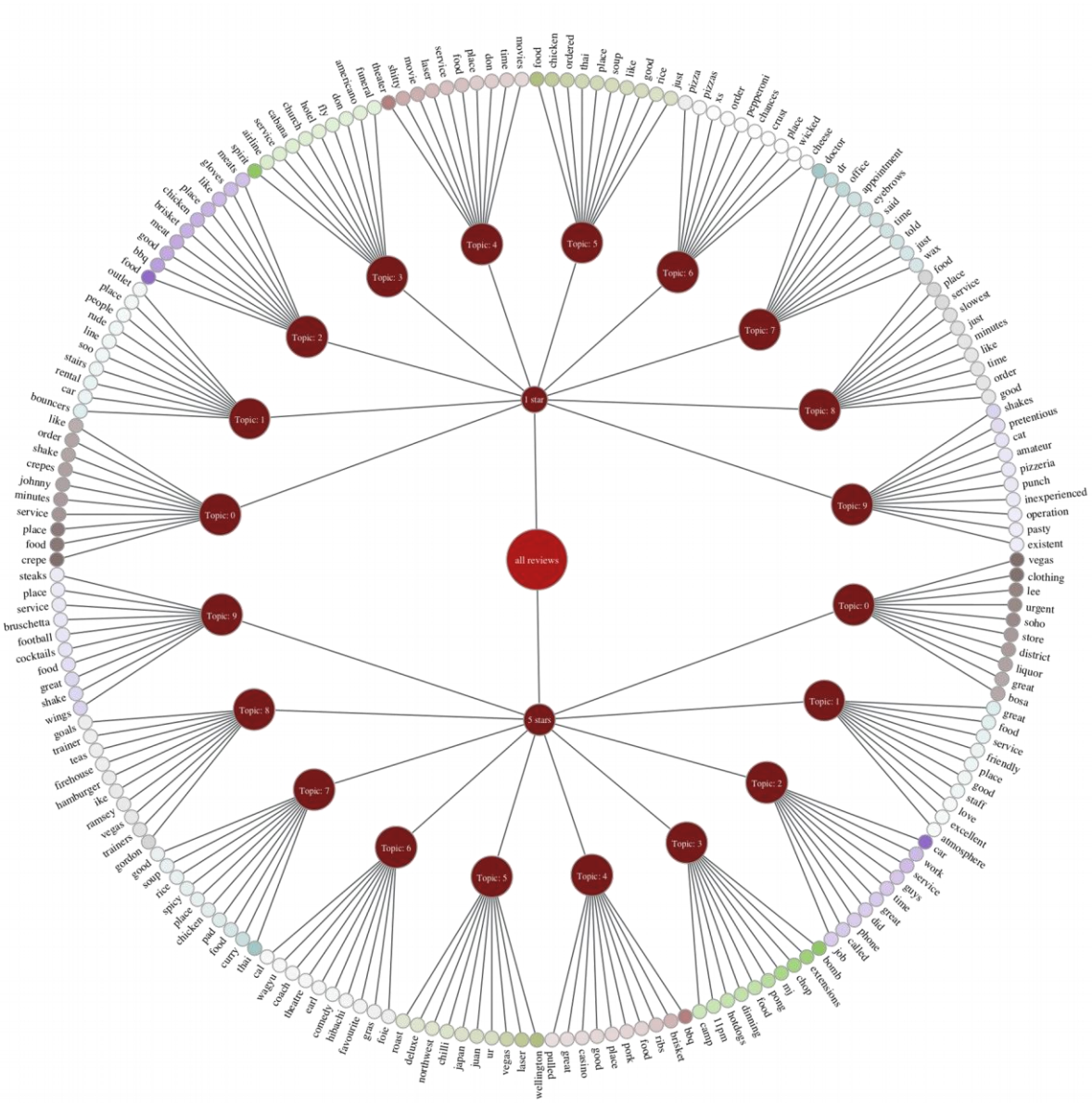


Probabilistic latent semantic analysis



Task 1.2

Comparison of reviews with ratings of one star and five stars



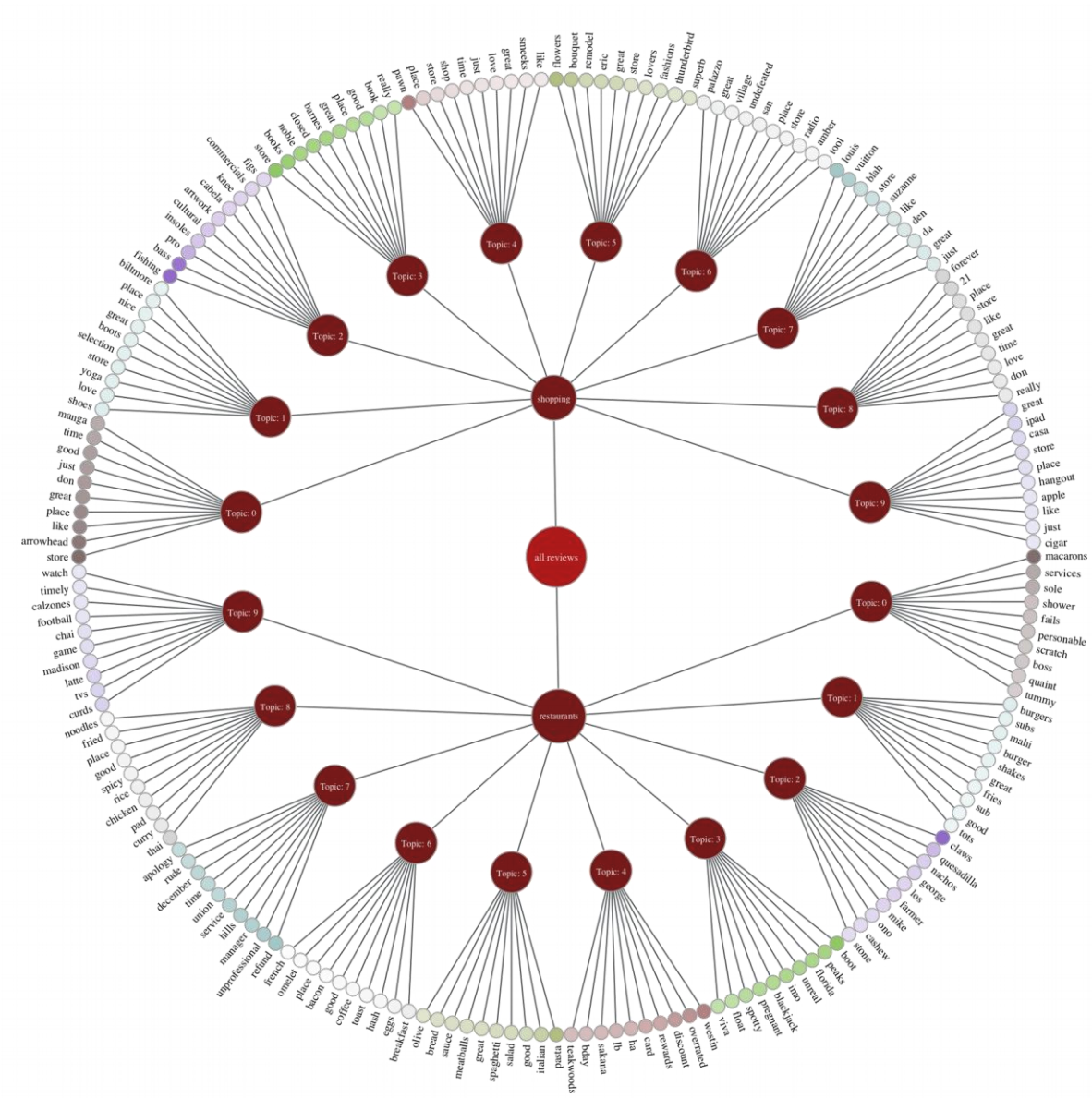
In principle on the basis of a visual analysis we can to say that the quality of topics is about the same for reviews with one star compared to five stars reviews.

The main difference that can be found is that words from five star reviews is a mix of general and positives adjectives, for instance, the topic #4 there is words about food like 'bbq', 'ribs' and at the same time there are 'great' or 'good'. From low rating reviews a good example of statement above could be Topic #9, food related words: 'shakes', 'pizzeria', 'pasty' and negative adjectives: 'inexperienced', 'pretentious'.

At the same time it is necessary to notice that low rated reviews still have positive adjectives as a part while we don't consider the same thing for high rated reviews.

An example of this can be Topic #2 from one star reviews, here we see restaurant related lexicon: 'bbq', 'meat' and positive words 'good', ‘like'.

Comparison of topics for restaurant reviews and doctor reviews



Now let's compare topics in both sub-groups of reviews. Topics for shopping reviews are quite clear distributed by type of shops. Here we can see book store topic, this is Topic #3 ('book', 'books', 'barnes', 'noble') or flower shops in Topic #5 ('flowers', 'bouquet').

In turn, the topics based on restaurant reviews are distributed by type of cuisine or type of restaurant. Here we can see topic related to Italian food (Topic #5: 'pasta', italian', 'meatballs', 'sauce', 'olive', 'bread') or tai cuisine (Topic #0: 'tai', 'carry', 'pad', 'chicken', 'rice', 'spicy', 'noodles').