Homework #3
CS 5665, Fall 2016

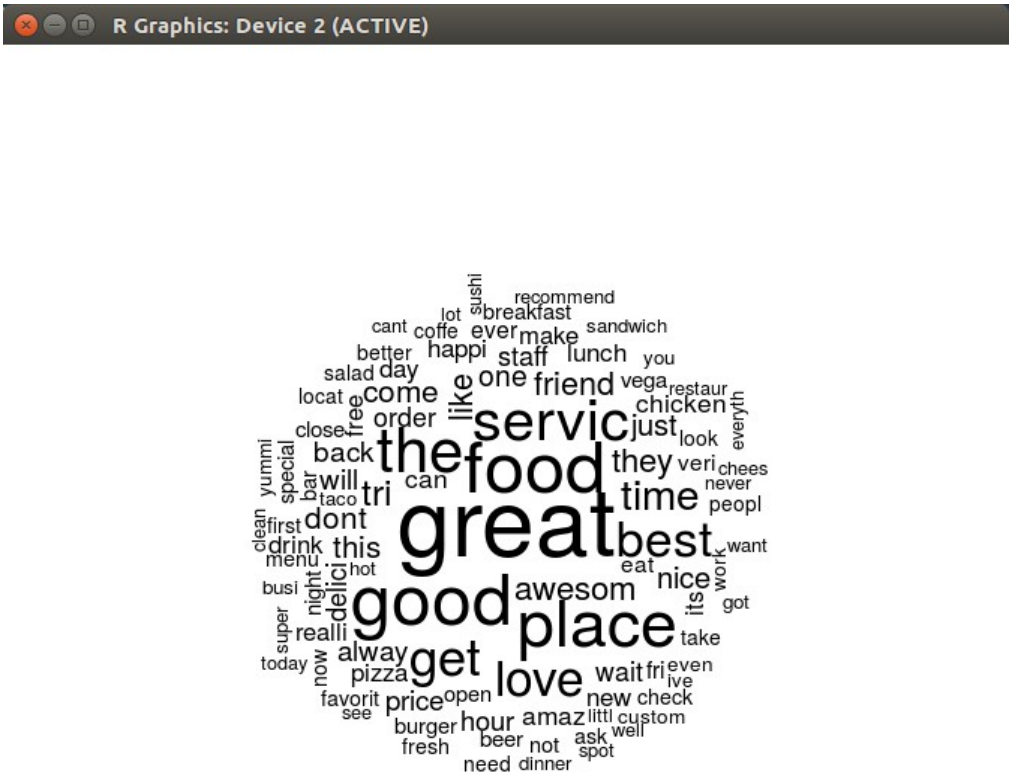
1) I choose Yelp Dataset from Yelp Challenge (https://www.yelp.com/dataset_challenge)

About Dataset:
+ 2.7M reviews and 649K tips by 687K users for 86K businesses
+ 566K business attributes, e.g., hours, parking availability, ambience.
+ Social network of 687K users for a total of 4.2M social edges.
+ Aggregated check-ins over time for each of the 86K businesses
+ 200,000 pictures from the included businesses

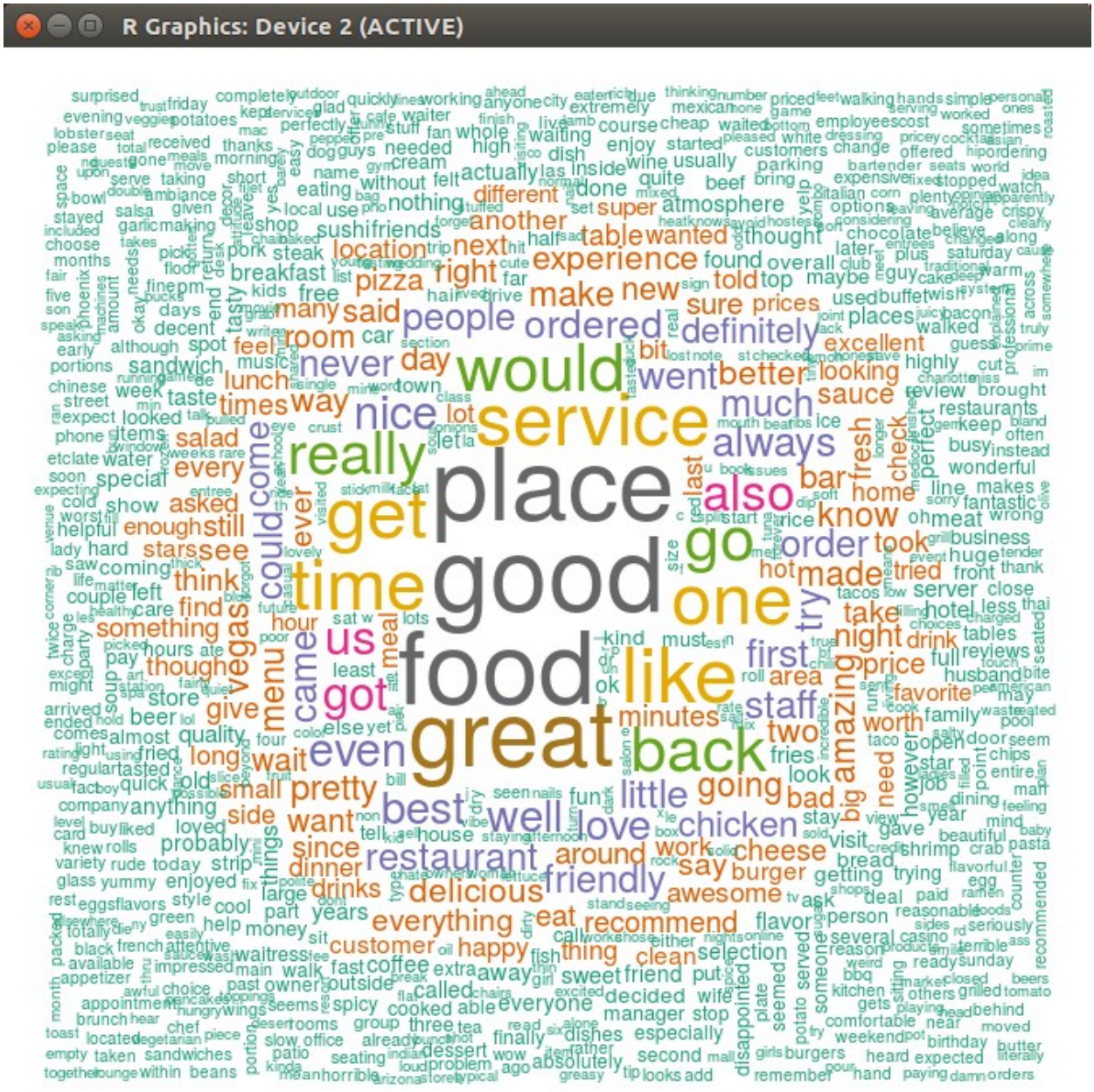
+ I was looking for what people are talking about most in there tips and reviews and compare.
+ I choose *word cloud for visualizing*, I tried two different ways of word cloud first only top 100 for tips and later no limits on the number of words and added colors based on the frequency.
+ Word cloud helps in visualizing what are the most frequent words in corpus of documents. The Shape, Color and Size varies with respect to the frequency, more frequent means large in size.
+ Observation of words used:

In tips people are mostly used words are Food, Service, Place, Great, Good, Love.
In Reviews people are mostly using words Good, Place, Food, Service, Restaurant, Friendly.

+ Below is a Word Cloud of 649K tips by 687K users for 86K businesses. (Top 100 most frequent)



+ Below is a Word Cloud of 2.7M reviews by 687K users for 86K businesses.



+ Reason I tried two different formation of word cloud is to compare the visualization structure. Colors and Size on the second WordCloud are choosen based on frequency of word.

+ For preprocessing, I am removing stop words and all non alphabetical letters from the text.

+ Downside of this visualization is, it conveys only TF of each word. In lot of situations it may not be very intuitive and helpful.

My Experience with 2.7M Yelp reviews

The size of the file with 2.7M reviews is 2.4GB. Question is how to process such large data ? I realized I can load it in memory and process, until I actually tried to load. Depending on system type, file will expand in RAM and in this case after loading file in memory there is no space left to perform processing. :(

(System configuration: i7-6700 3.3Gh 8 cores and RAM 16 GB DDR4-2200, it is one of the latest configuration you can get for a desktop. Thanks to my advisor Prof. Lee :))

+ First, I tried cleaning using OpenRefine but it will throw runtime JAVA exception after ~65326 number of rows (it seems they use 32 bit int value for integer as row count: Definitely a bug ;))

+ I tried using Tableau, no matter how much they claim for being able to handle big data. My system hanged atleast 5 times with no success.

+ Then I decided to load file in memory using Python, success in loading file but no more memory space for any processing.

+ I got slight success with R, loading file and processing it one step, later again out of memory + Swap memory(4GB). (total memory used 14+4 = 18GB) and I could only achieve one step of processing.

+ Then I load the data on Mongo using mongoimport. Finally success and I was able to atleast see my *data*, *how it looks like*. :)

+ But how to visualize data :(Mongo has no visualization tool.

+ Next I tried reducing my data, using mongoexport took text out of data and dumped it in a CSV file leaves me with 1.7GB (shredding 0.7GB).

+ The only way I could think to process such data was to distribute computation. I used Mapper from Python which is a built in function for distributed computing (similar to Map-Reduce). It uses multi threading to distribute computation.

+ Without distributed computation, with my calculation it was taking about ~7 hours to process the data, but using it took me ~126 Sec. (Even though I have 8 cores it cannot be thought as time should reduce by 8 times, because the way Mapper process data will reduce the computation cost exponentially.)

+ Finally, I perform word count (TF) in Mapper and build a dictionary and store it in a file sized 5.6MB. (From 1.7GB) :)

+ Resultant was the second Word Cloud of all reviews. It took me 2 days to build a word cloud.