CS598 DMC Task 1

Exploration of the Dataset

Yanislav Shterev (shterev2@illinois.edu, netid shterev2)

Overview:

The purpose of the following document is to analyze part of Yelp's reviews academic dataset and give an initial idea on what the entity-relationships and mine this data set to discover interesting and useful knowledge. To accomplish the visualizations below I have used multiple topic models, Python libraries like **NLTK**, **sklearn**, **genism models**, **matplotlib and graphlab**(the non-commercial license) for topic modeling and text processing. **D3.js** was used to generate the Radial Dendrograms and **matplotlib.pyplot** and Python's **WordCloud** to plot the word cloud images.

1. **Task 1.1: Application of a topic model**

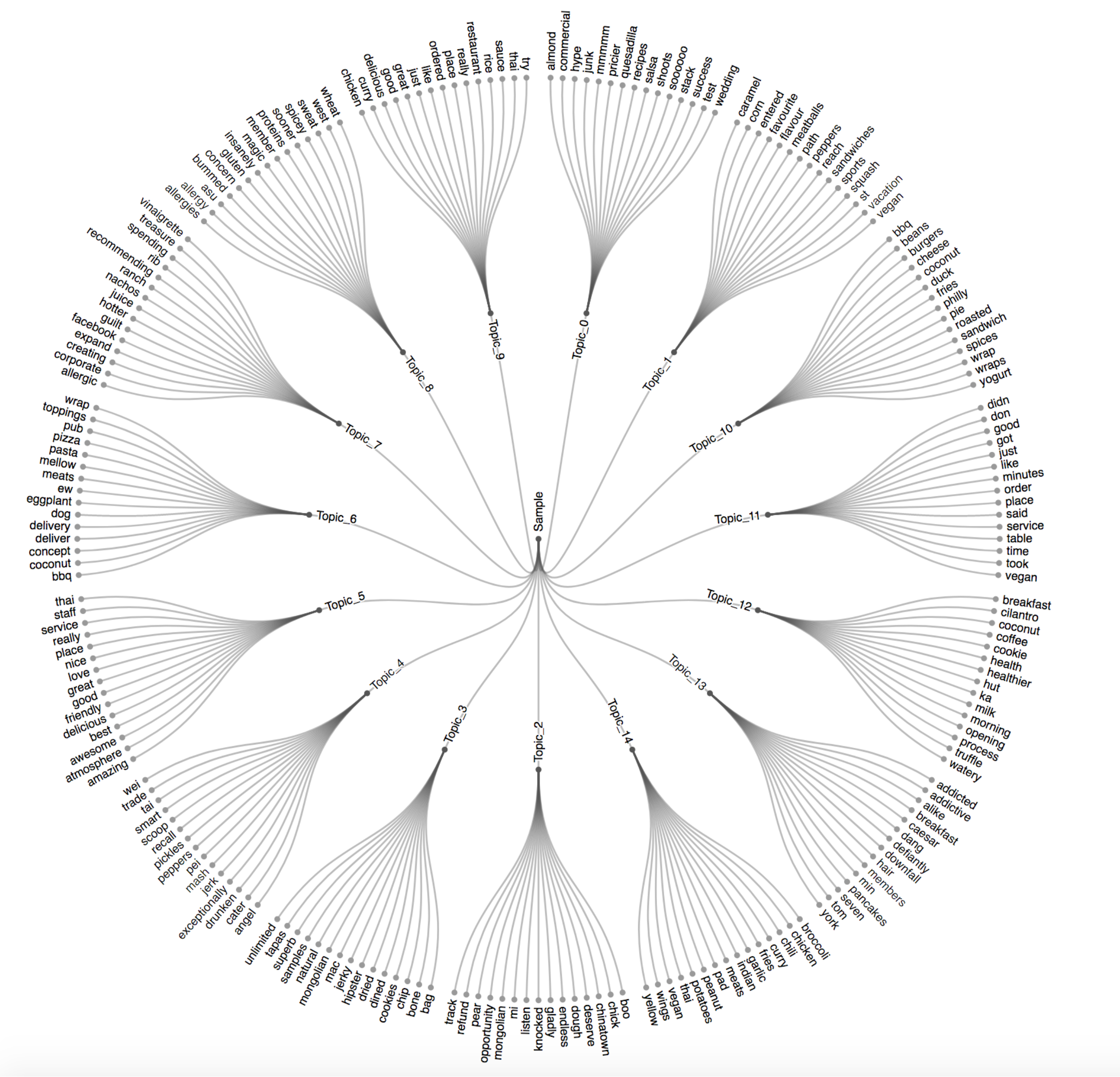
Topic models used:
- **Collapsed Gibbs Sampling** (https://en.wikipedia.org/wiki/Gibbs_sampling) is a Markov Chain Monte Carlo algorithm which internally uses multivariate probability distribution. Parameters used – number of iterations being 200 and number of topics being 10.

-**LDA** - The main idea of the LDA model assumes that each document may be viewed as a mixture of various topics, where a topic is represented as a multinomial probability distribution over words. Parameters used – number of clusters(topics) being 15 and transformation of the initial reviews into a sparse corpus using the genism matutils library.

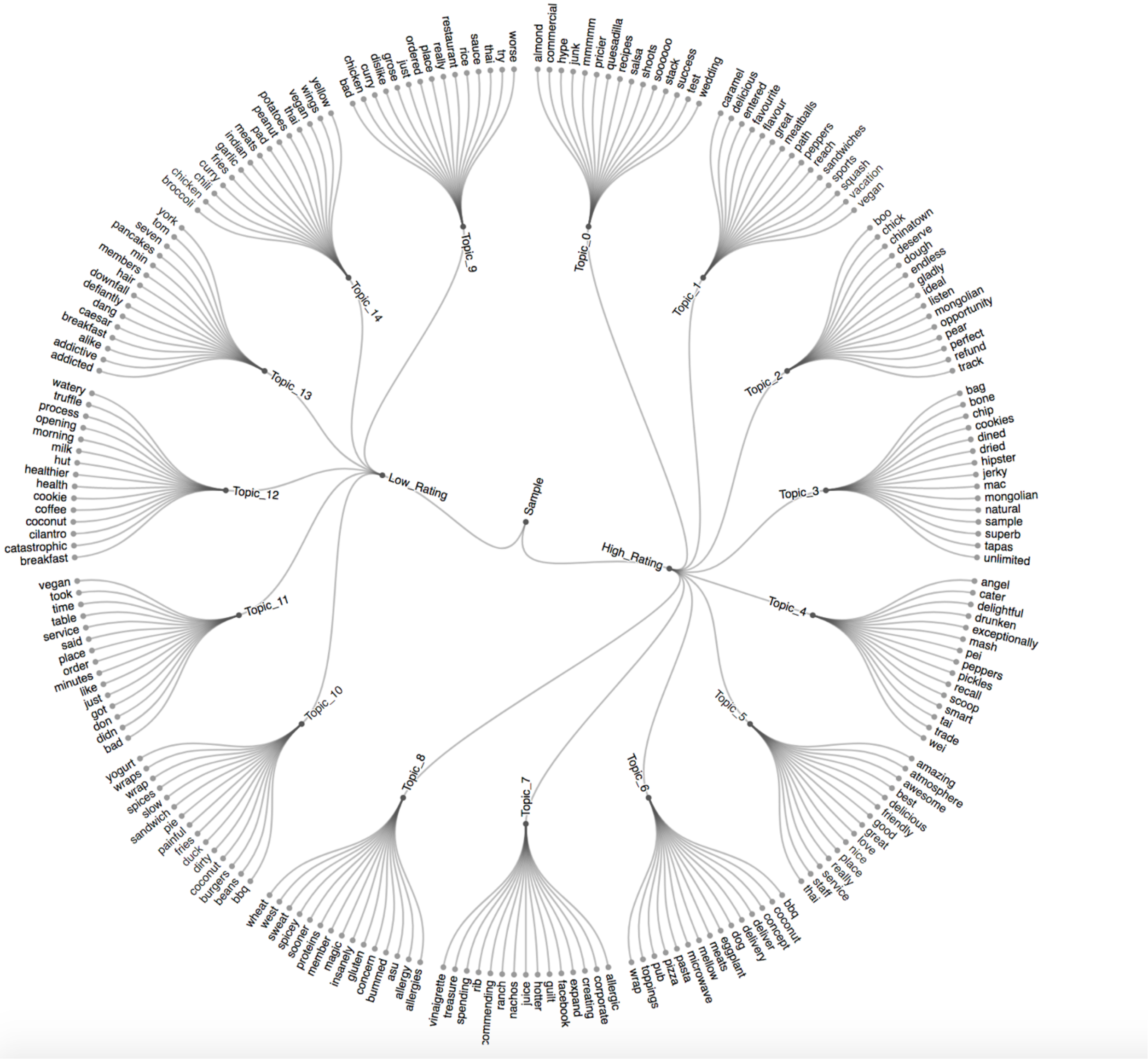
Even though both models are probabilistic the main difference is that the Gibbs sampling is optimized to run for large amounts of data and has improved performance. Another great feature of Gibbs sampling is that we are able to find correlation between nearby words due to the Markov chain algorithm. This gives us more accurate results while doing topic mining.

2. **Task 1.1: Generated visualization**

a) The first Radial Dendrogram illustrates the top 15 topics that were produced by Latent Dirichlet allocation and the most popular words among these topics.

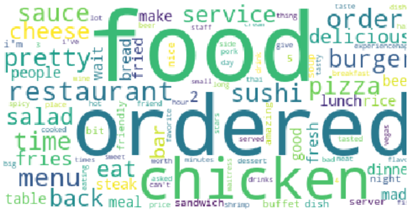


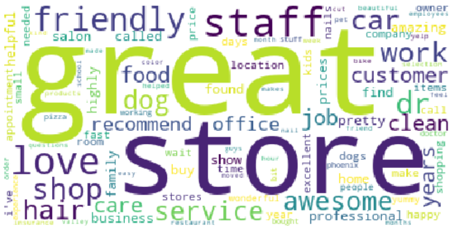
b) The second Radial Dendrogram presents the top topics in two categories- low and high rating. In order to do that, I divided the dataset into two subsets. One containing reviews with 1 and 2 stars and another having high rating reviews – 5 stars. As part of the analysis we can clearly notice some of the most frequent words in these topics are with negative meaning for low rating topics and positive for high rating topics. For instance, words like 'dislike, grouse, worse' in Topic_9 clearly emphasize on low rating reviews. This means the same approach can be used for sentiment analysis.

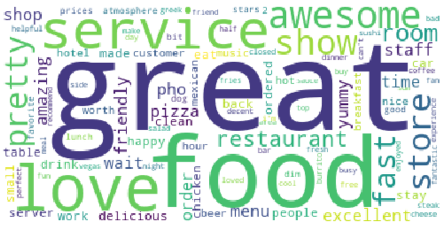

Sample
High_Rating
Low_Rating
Topic_0
almond
commercial
hype
junk
mmmmm
pricier
quesadilla
recipes
salsa
shoots
soooooo
stack
success
test
wedding
Topic_1
caramel
delicious
entered
favourite
flavour
great
meatballs
path
peppers
reach
sandwiches
sports
squash
vacation
vegan
Topic_2
boo
chick
chinatown
deserve
dough
endless
gladly
ideal
listen
mongolian
opportunity
pear
perfect
refund
track
Topic_3
bag
bone
chip
cookies
dined
dried
hipster
jerky
mac
mongolian
natural
sample
superb
tapas
unlimited
Topic_4
angel
cater
delightful
drunken
exceptionally
mash
pei
peppers
pickles
recall
scoop
smart
tai
trade
wei
Topic_5
amazing
atmosphere
awesome
best
delicious
friendly
good
great
love
nice
place
really
service
staff
thai
Topic_6
bbq
coconut
concept
deliver
delivery
dog
eggplant
meats
mellow
microwave
pasta
pizza
pub
toppings
wrap
Topic_7
allergic
corporate
creating
expand
facebook
guilt
hotter
juice
nachos
ranch
commending
rib
spending
treasure
vinaigrette
Topic_8
allergies
allergy
asu
bummed
concern
gluten
insanely
magic
member
proteins
sooner
spicey
sweat
west
wheat
Topic_10
bbq
beans
burgers
coconut
dirty
duck
fries
painful
pie
sandwich
slow
spices
wrap
wraps
yogurt
Topic_11
bad
didn
don
got
just
like
minutes
order
place
said
service
table
time
took
vegan
Topic_12
breakfast
catastrophic
cilantro
coconut
coffee
cookie
health
healthier
hut
milk
morning
opening
process
truffle
watery
Topic_13
addicted
addictive
alike
breakfast
caesar
dang
defiantly
downfall
hair
members
min
pancakes
seven
tom
york
Topic_14
broccoli
chicken
chili
curry
fries
garlic
indian
meats
pad
peanut
potatoes
thai
vegan
wings
yellow
Topic_9
bad
chicken
curry
dislike
gross
just
ordered
place
really
restaurant
rice
sauce
thai
try
worse

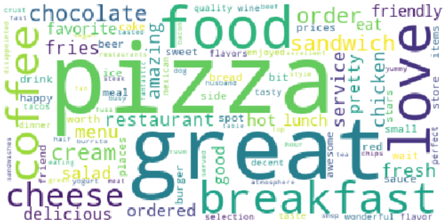
c) A third type of visualizations based on multiple subsets were created from all the reviews (not only restaurants). The entire subset was again divided into positive (5 star) reviews and negative (1 or 2 star) reviews.

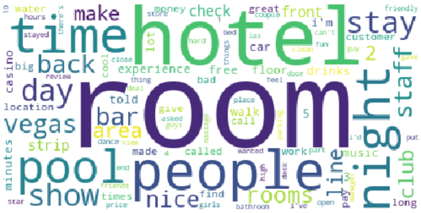
Positive Word Cloud based on different topics:



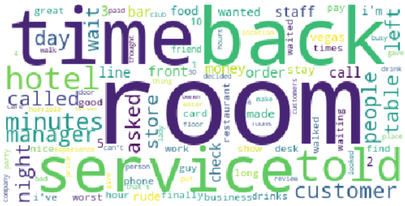


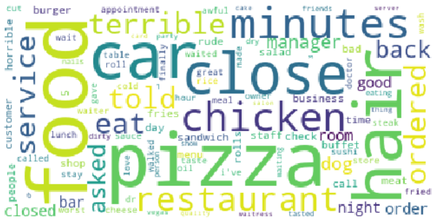


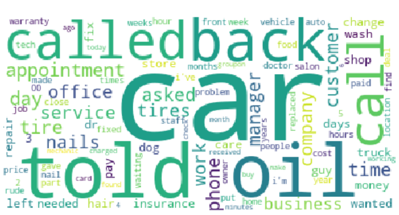


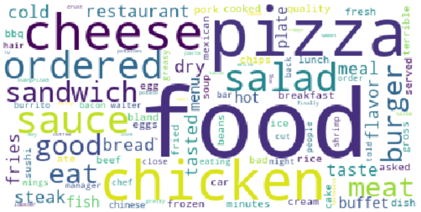


Negative word cloud topics:









3. **Task 1.2: Generated sets of topics**

Data Subsets:

- The data used for generating the dendrograms was only extracted from reviews for restaurants.
- Subsets of negative and positive restaurant reviews were also extracted and visualized.
- All the reviews were taken in order to generate the word clouds.
- The forth group of subsets were generated from all the reviews and divided into positive and negative subsets.

4. **Task 1.2: Visualization of comparison**

Transformations that were used are: representing the text as **bag of words** and running word tokenizer. Then word stemming was applied and word lemmatizer was used to group together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meaning to one word. Libraries that were used are the **WordLemmatizer** and **PorterStemmer** from **NLTK**.

The visualizations based on different subsets and different topic models are self-explanatory and show exactly what the narrator wants to explain. The radial dendrograms are aimed to show the breakdown of different topics which are popular among the Yelp reviews. They also clearly show what the main words in these topics are and how they are correlated to one another. For instance, we can see group of words wrapped in topic 8 which are all talking about food ingredients and possible allergies.

The second dendrogram breaks down to more granular subsets and models their topics. It is used to dive deeper and show some correlation between negative or positive reviews and the most frequently used words there.

Word Cloud was used as third visualization and was applied to multiple different subsets: All the reviews (not only the restaurants like on the Dendrograms), splitting all the reviews into positive and negative and analyzing a specific cosine.

All the code and extracted data can be found on: https://github.com/yan6pz/CS-598-Data-Mining-Capstone