**CS598 DMC Task 1**

Exploration of the Dataset

Yanislav Shterev ([shterev2@illinois.edu](mailto: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.

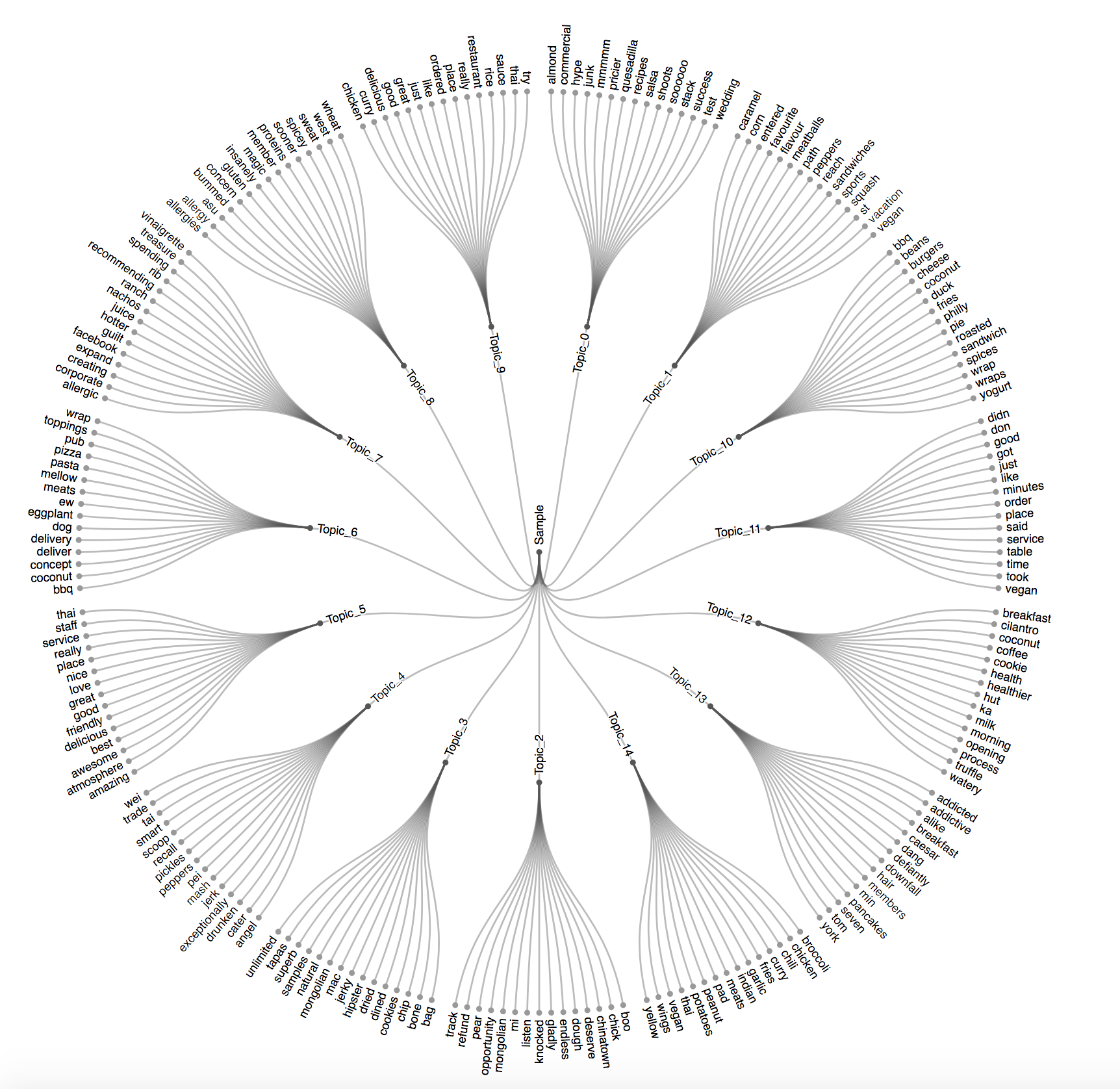
1. **Task 1.1: Application of a topic model**

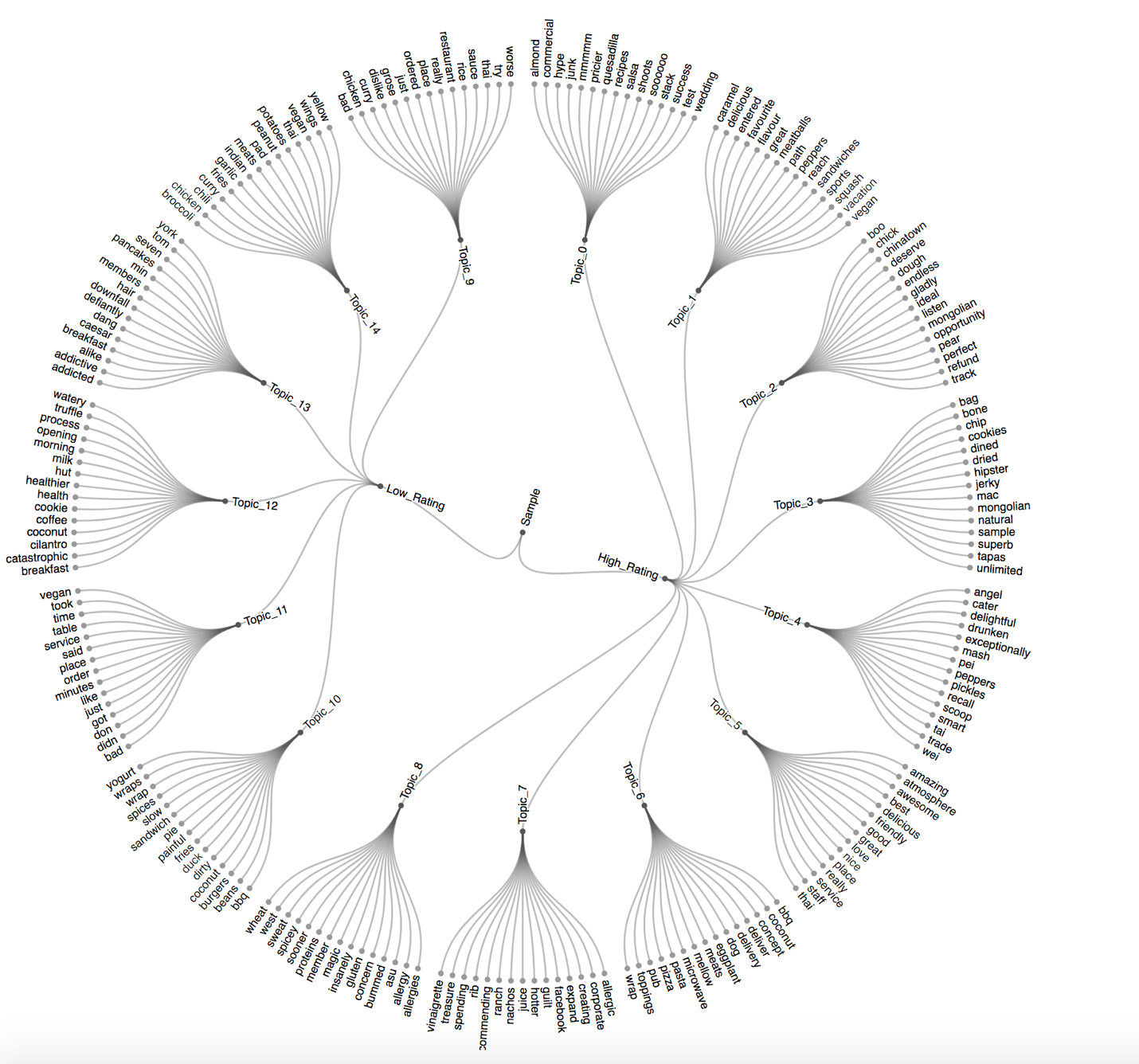
Topic models used:

- 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.

1. **Task 1.1: Generated visualization**
2. The first Radial Dendrogram illustrates the top 15 topics that were produced by Latent Dirichlet allocation and the most popular words among these topics.
3. 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.



Positive Word Cloud



Negative word cloud topics:



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

Data Subsets:

The data used for generating the dendrograms was only extracted from reviews for restaurants.

However, a wider reviews sets were taken in order to generate the word clouds. As noted

1. **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.

World cloud- additional stop words removal

1. Your opinions about whether the results you generated make sense or are useful in any way. Your description should be detailed enough to allow others to replicate your work. Are there any particular aspects of your visualization to which you would like to bring attention? What do you think the data and your visualization show?