STAT 628 Module 2 Presentation-1

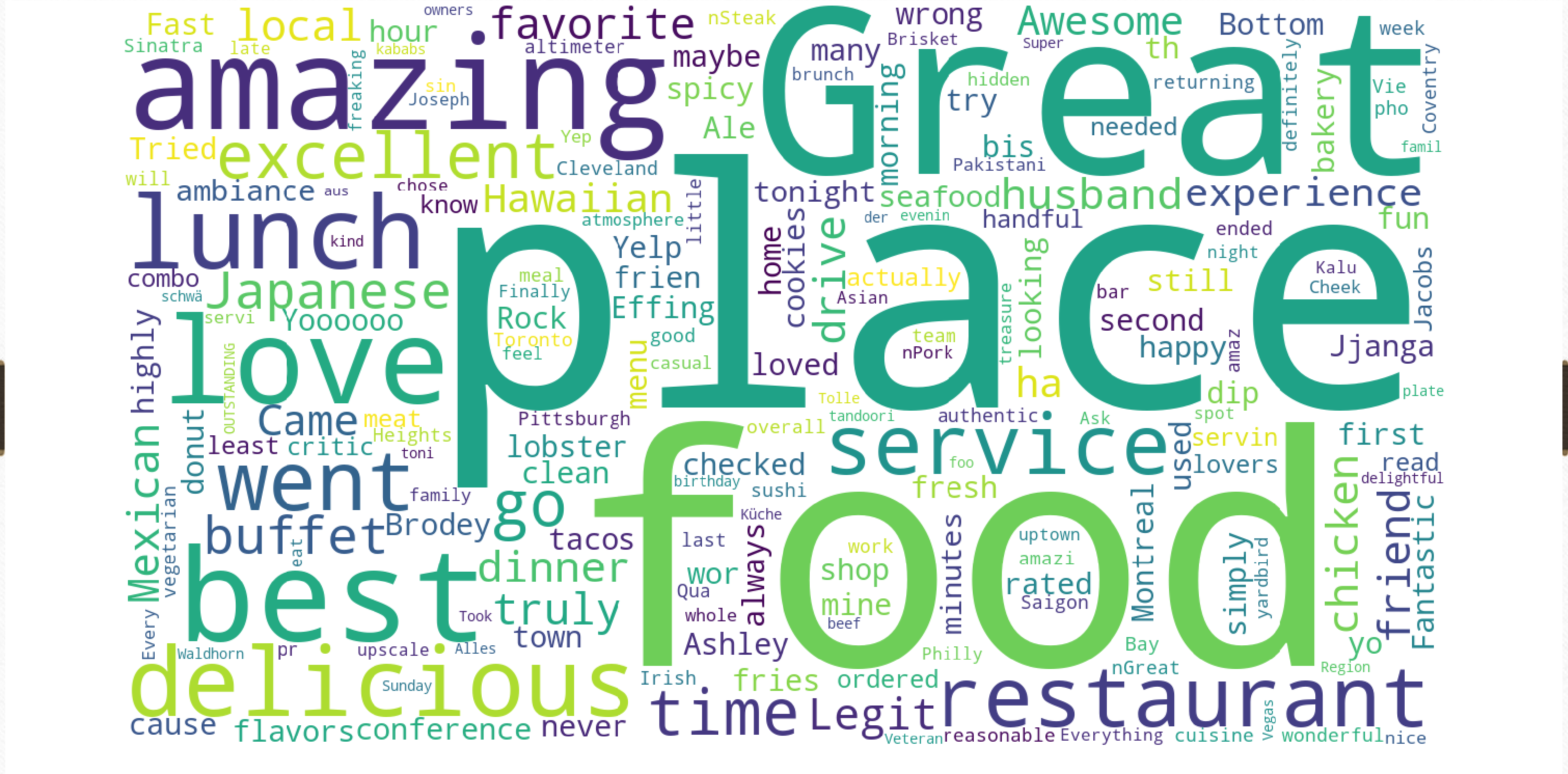
Group 6: Qizheng Ren, Hanmo Li, Lixia Yi, Jiacheng Xu

WHAT TO VECTOR

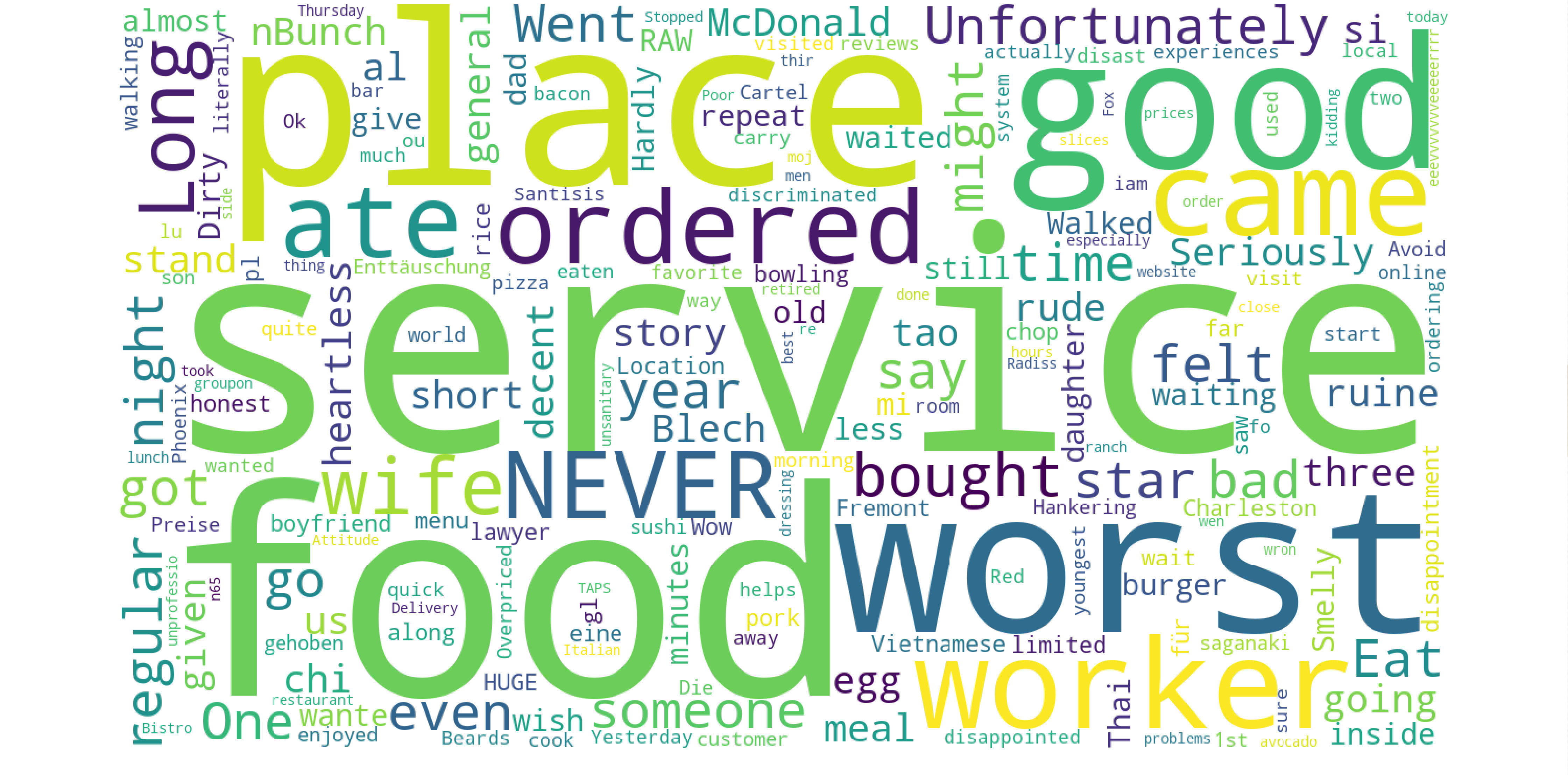
Data Overview

- General Idea: Word frequency in reviews differ
- Example:

    5 stars and 1 star reviews have different word frequencies



Word Cloud of 5 Star Reviews



Word Cloud of 1 Star Reviews

HOW TO VECTOR

Extract information from reviews

- Bag of Words
  - Sparse matrix of word frequency

Extract information from reviews

- Dimension reduction
  - Word2Vec
  - K-means
- Reduce dimension from 210,000 to 1000

Extract information from reviews

- Self-defined emotion score
  - Multiply lexicon scores with the weights from TF-IDF
  - positive and negative degree of reviews

Factor Variables

- There are 6 factor variables which maybe be useful
  - “name”, “time”, “city”, “longitude”, “latitude” and “category”
- Make them interpretable for machine to fit model
- Transform them into (0,1) matrices

Category to Sparse Matrix

- Under “Restaurant” category, there are 154 different subcategories
- There is not only “Restaurant”
- “Restaurant” tag it won’t be included in (0,1) matrix
- Then we create a category matrix with 154 columns
- Example: **['Burgers', 'Fast Food', 'Restaurants’]**

Afghan	……	Burgers	……	Fast Food	……
0	0……	1	0……	1	0……

Date & Location to Sparse Matrix

- Similar transformation for date & location data
- Matrix columns are year (2005-2017), month (1-12), day (1-31), week (1-53)
- Example: 2017-01-15

2005	……	2017	01	……	12	1	……	15	……	31	1	……	3	……	53
0	0…	1	1	0…	0	0	0…	1	0…	0	0	0…	1	0…	0

Year Month Day Week

Date & Location to Sparse Matrix

- Unique longitude and latitude determine name and city for that restaurant
- 43035 different locations, encode them from 0 to 43034
- Example

index	stars	name	text	date	city	longitude	latitude	category
1	1	McDonald's	……	……	Glendale	-112.205020	33.509597	……
2	1	McDonald's	……	……	Las Vegas	-115.256458	36.181713	……

index	0	1	……
1	1	0	0……
2	0	1	0……

VECTOR TO WHAT

Analysis Plan

- Feature Selection
- Model Selection
- Parameters Tuning
- "Deep Learning"
- …

Step1: Feature Selection

- Feature importance
- Feature combination
  - Among reviews; time; categories...

Step2: Model Selection

- Linear Regression
- Support Vector Machine
- Random Forest
- eXtreme Gradient Boosting (XGBoost)
- ...

Step3: Parameters Tuning

- Random forest
  - max depth, number of trees...
- Linear Regression
  - L1&L2 regularization...

Step4: "Deep Learning"

- Fetch deep patterns
  - dependent syntax; set phrase

Thank you!