Yelp Review Analysis

Data Visualization mid presentation

Dataset Intro

- Yelp, Inc. is an American company that develops the Yelp.com website and the Yelp mobile app, which publish **crowd-sourced reviews** about businesses.

- The Yelp dataset is a subset of our **businesses, reviews, and user data** for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

business.json

Contains business data including location data, attributes, and categories.

review.json

Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

checkin.json

Checkins on a business.

user.json

User data including the user's friend mapping and all the metadata associated with the user.

tip.json

Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

photo.json

Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

Dataset Intro

YELP, INC. AND 1 COLLABORATOR · UPDATED 2 MONTHS AGO

kaggle

Yelp Dataset

A trove of reviews, businesses, users, tips, and check-in data!

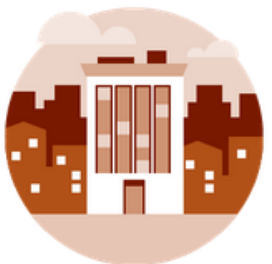
Workflow tips:
1.Kaggle kernel
2.filters, constraints, limitations

The Dataset

6,990,280 reviews

150,346 businesses

200,100 pictures



11 metropolitan areas

908,915 tips by 1,987,897 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 131,930 businesses

Dataset components

Blue: Identifying
Orange: Information
Purple: Filter

business.json

Contains business data including location data, attributes, and categories.

```
{
    // string, 22 character unique string business id
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // string, the business's name
    "name": "Garaje",

    // string, the full address of the business
    "address": "475 3rd St",

    // string, the city
    "city": "San Francisco",

    // string, 2 character state code, if applicable
    "state": "CA",

    // string, the postal code
    "postal code": "94107",

    // float, latitude
    "latitude": 37.7817529521,

    // float, longitude
    "longitude": -122.39612197,

    // float, star rating, rounded to half-stars
    "stars": 4.5,

    // integer, number of reviews
    "review_count": 1198,

    // integer, 0 or 1 for closed or open, respectively
    "is_open": 1,

    // object, business attributes to values. note: some attribute values m
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },

    // an array of strings of business categories
    "categories": [
        "Mexican",
        "Burgers",
        "Gastropubs"
    ],

    // an object of key day to value hours, hours are using a 24hr clock
    "hours": {
        "Monday": "10:00-21:00",
        "Tuesday": "10:00-21:00",
        "Friday": "10:00-21:00",
        "Wednesday": "10:00-21:00",
        "Thursday": "10:00-21:00",
        "Sunday": "11:00-18:00",
        "Saturday": "10:00-21:00"
    }
}
```

Dataset components

review.json

Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

```
{
    // string, 22 character unique review id
    "review_id": "zdSx_SD6obEhz9VrW9uAWA",

    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

    // integer, star rating
    "stars": 4,

    // string, date formatted YYYY-MM-DD
    "date": "2016-03-09",

    // string, the review itself
    "text": "Great place to hang out after work: the prices are decent, and

    // integer, number of useful votes received
    "useful": 0,

    // integer, number of funny votes received
    "funny": 0,

    // integer, number of cool votes received
    "cool": 0
}
```

Special: Time-related

Blue: Identifying
Orange: Information
Purple: Filter

user.json

User data including the user's friend mapping and all the metadata associated with the user.

```
{
    // string, 22 character unique user id, maps to the user in user.json
    "user_id": "Ha3iJu77CxlrFm-vQRs_8g",

    // string, the user's first name
    "name": "Sebastien",

    // integer, the number of reviews they've written
    "review_count": 56,

    // string, when the user joined Yelp, formatted like YYYY-MM-DD
    "yelping_since": "2011-01-01",

    // array of strings, an array of the user's friend as user_ids
    "friends": [
        "wqoXYLWmpkEH0YvTmHBsJQ",
        "KUXLLiJGrjtSsapmxmpvTA",
        "6e9rJKQC3n0RSKyHLViL-Q"
    ],

    // integer, number of useful votes sent by the user
    "useful": 21,

    // integer, number of funny votes sent by the user
    "funny": 88,

    // integer, number of cool votes sent by the user
    "cool": 15,

    // integer, number of fans the user has
    "fans": 1032,

    // array of integers, the years the user was elite
    "elite": [
        2012,
        2013
    ],
```

Dataset components

tip.json

Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

```
// string, text of the tip
"text": "Secret menu - fried chicken sando is da bombbbbbb Their zapatos

// string, when the tip was written, formatted like YYYY-MM-DD
"date": "2013-09-20",

// integer, how many compliments it has
"compliment_count": 172,

// string, 22 character business id, maps to business in business.json
"business_id": "tnhfDv5Il8EaGSXZGiuQGg",

// string, 22 character unique user id, maps to the user in user.json
"user_id": "49JhAJh8vSQ-vM4Aourl0g"
```

checkin.json

Checkins on a business.

```
{
    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg"

    // string which is a comma-separated list of timestamps for each checki
    "date": "2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016-10-15 02:45:18,
}
```

photo.json

Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

```
{
    // string, 22 character unique photo id
    "photo_id": "_nN_DhLXkfwEkwPNxne9hw",
    // string, 22 character business id, maps to business in business.json
    "business_id" : "tnhfDv5Il8EaGSXZGiuQGg",
    // string, the photo caption, if any
    "caption" : "carne asada fries",
    // string, the category the photo belongs to, if any
    "label" : "food"
}
```

Task

我们可以提出许多问题：
评论模式在时间和地理上有什么不同，评论和企业的特点是什么，如何通过评论有效地提供企业的可视化摘要？

Analysis
1.The city/area's **business components: Industry-category-style**
2.Single business's **review dashboard**
3.**Interactive map** business exploring (optional)

Visual Design

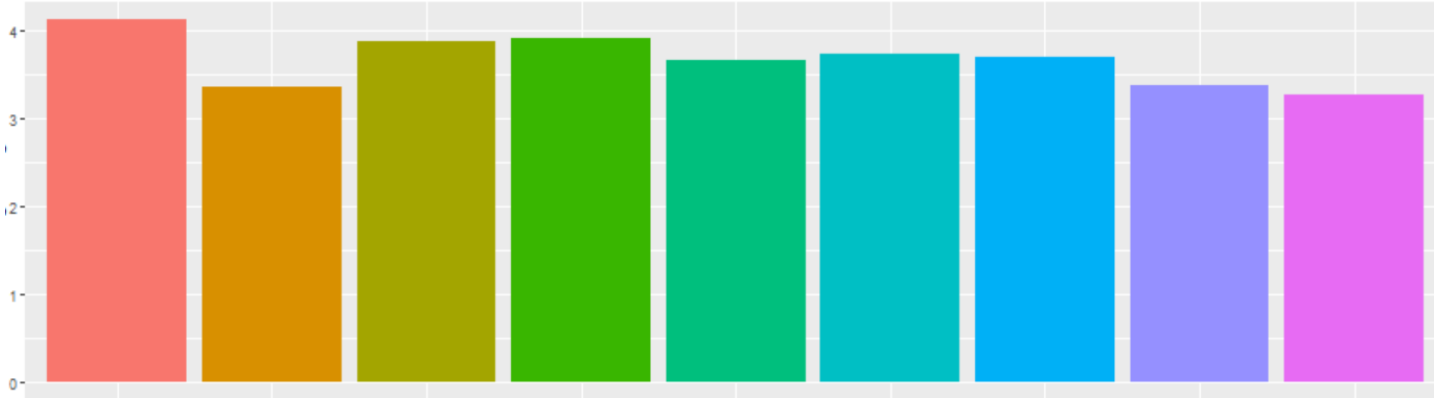
Analysis

The city/area's business components: Industry-category-style

Ratings – Scale

City	Business_name	Address	Postal_code	star_ratings	review_count	Category
Toronto	A Firkin Pub	3335 Bloor Street W	85016	3	5	Nightclub

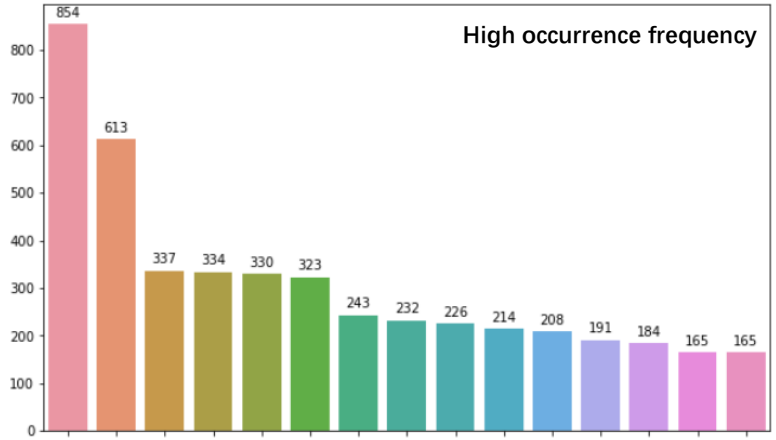
Average ratings/Business Scale



Active life Repairs Beauty Entertainment ……

City: Toronto

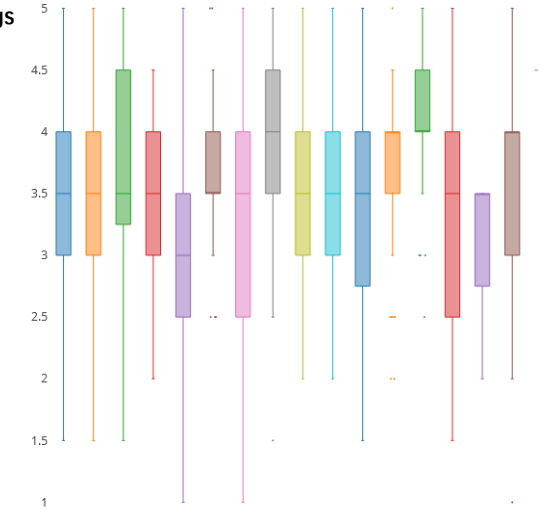
Time scale：1 week/month/year



City: Toronto

Category: Food

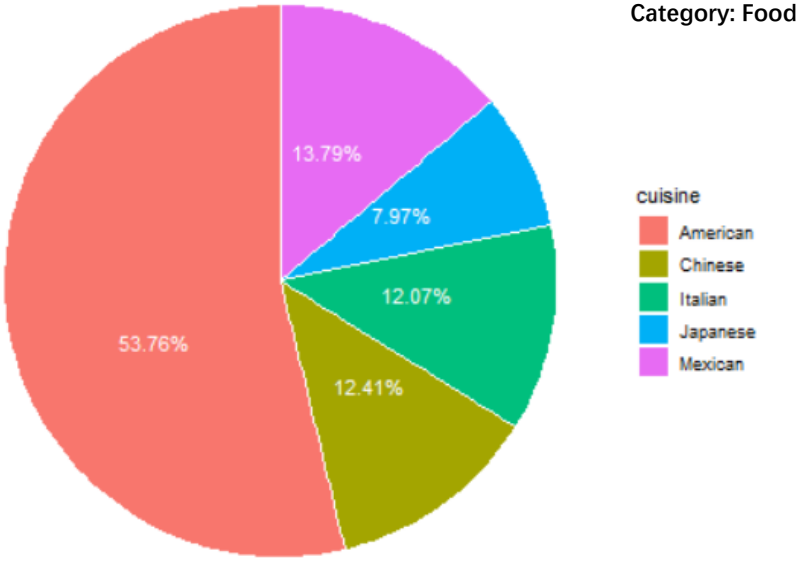
Time scale：1 week/month/year



Industry/category/style

City: Toronto

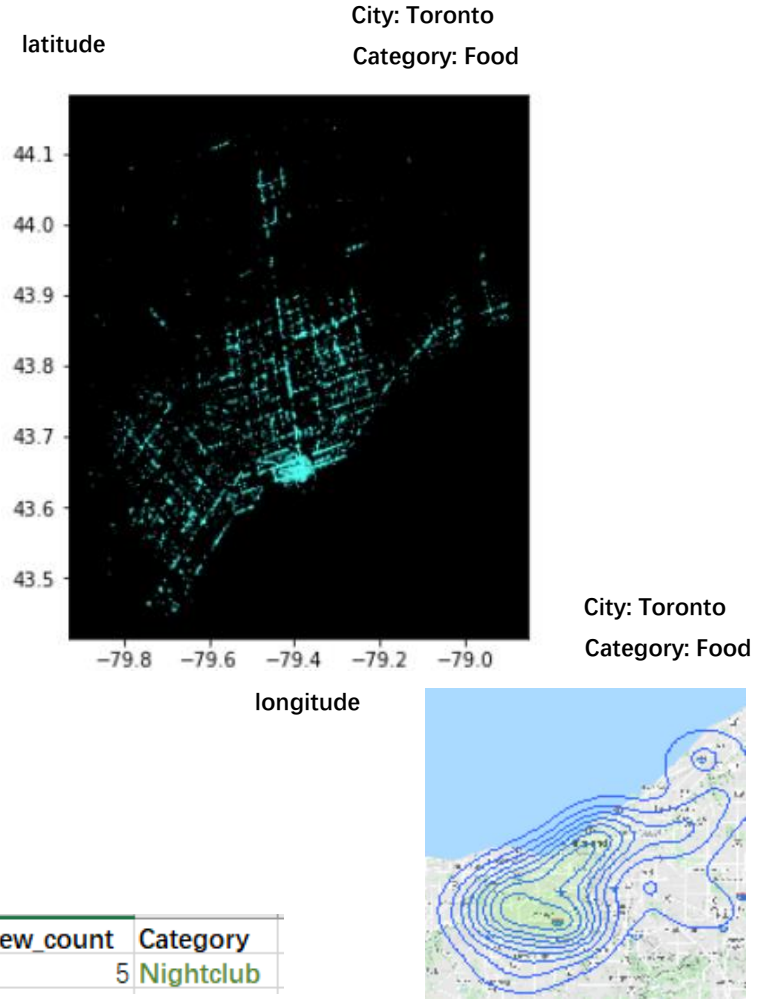
Category: Food



Visual Design

Introduce Google map API?
interactive visualization map

Otherwise: Heat map



City	Business_name	Address	Postal_code	star_ratings	review_count	Category
Toronto	A Firkin Pub	3335 Bloor Street W	85016	3	5	Nightclub

Visual Design

Single business's Review Dashboard

Name: Ma's Noodle City: Toronto Category: Food Key: Fast food, Chinese food

Recent Highly Voted Review

review1

review2

review3

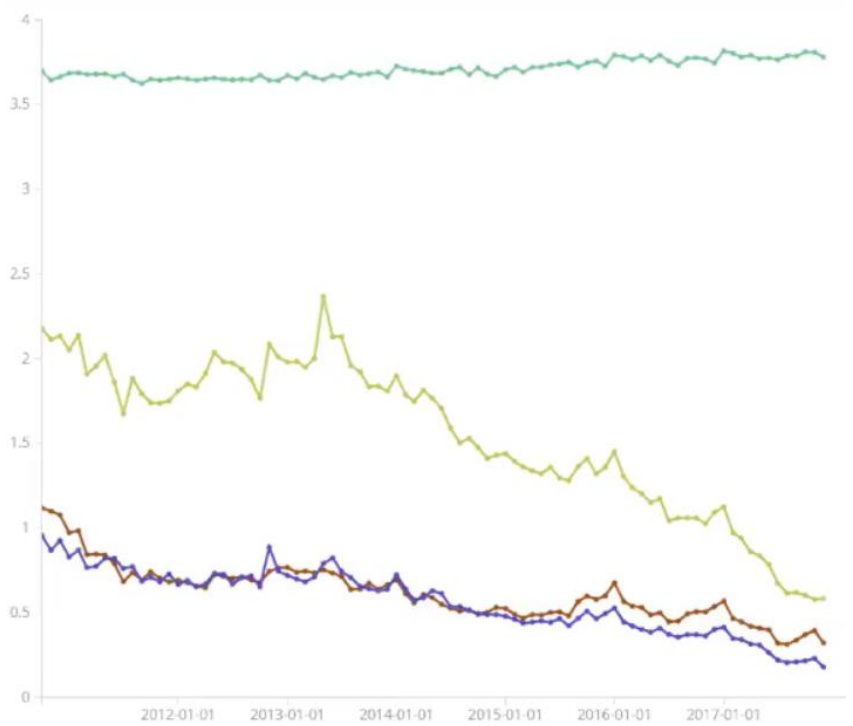
Time scale：1 week/month/year

Recent tips

tip1

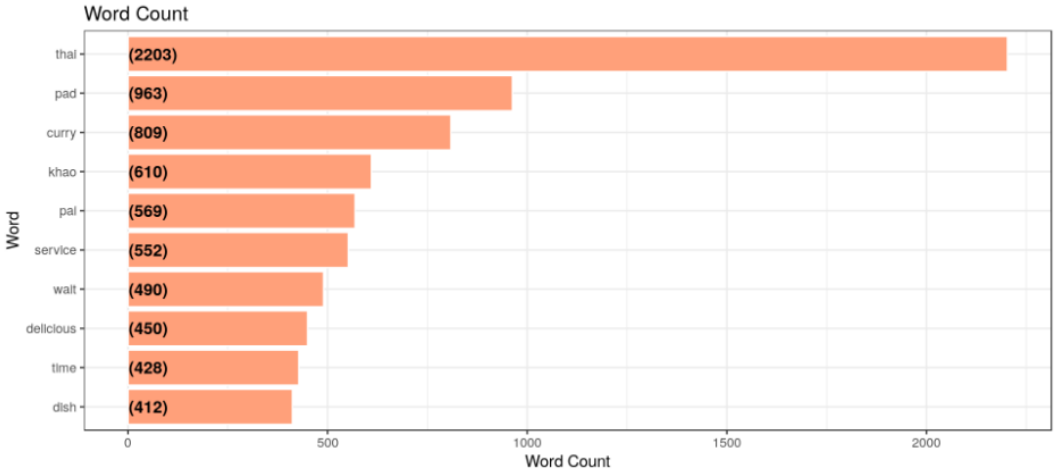
tip2

Ratings/review number



Time

Review word count



Tips word cloud

Time scale：1 week/month/year

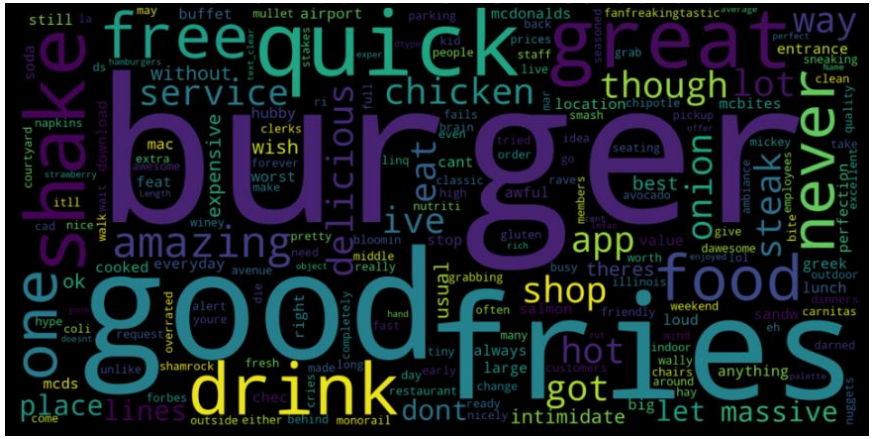
Review word cloud

Time scale：1 week/month/year

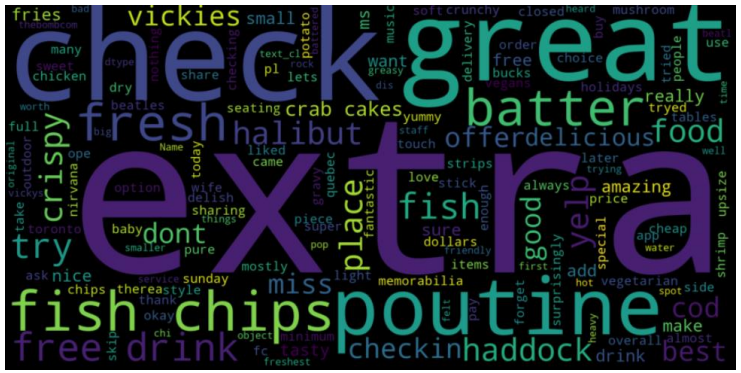
Visual Design

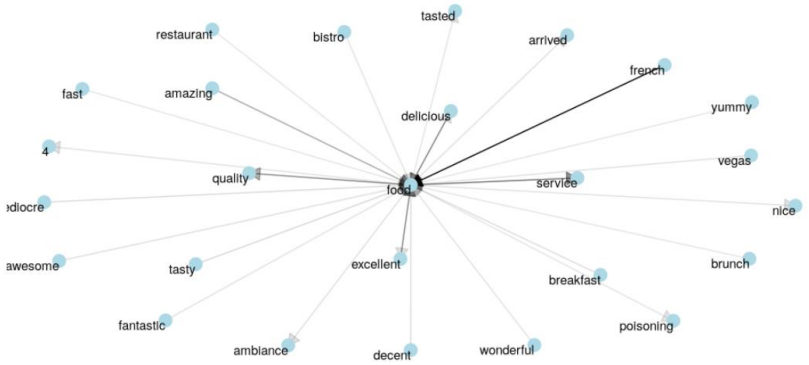
Single business's **Review Dashboard**

Top10 Average / Category Average word cloud



Necessary addition:
Extra what?
Extra sit, extra parking space?

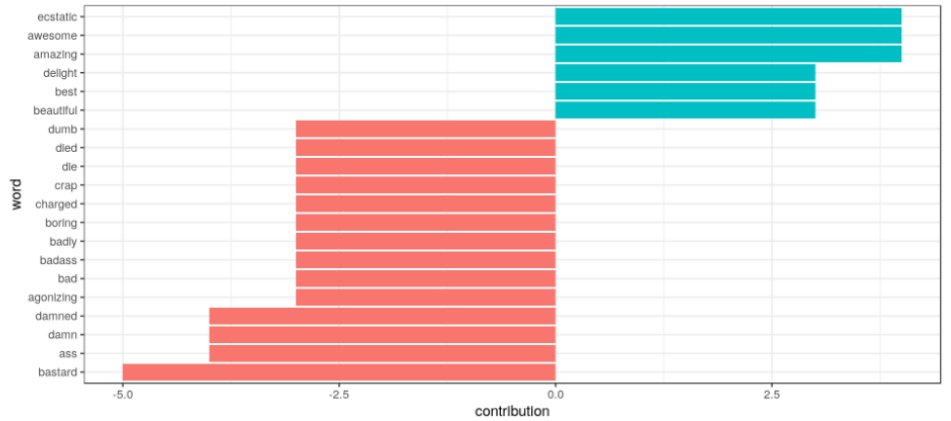




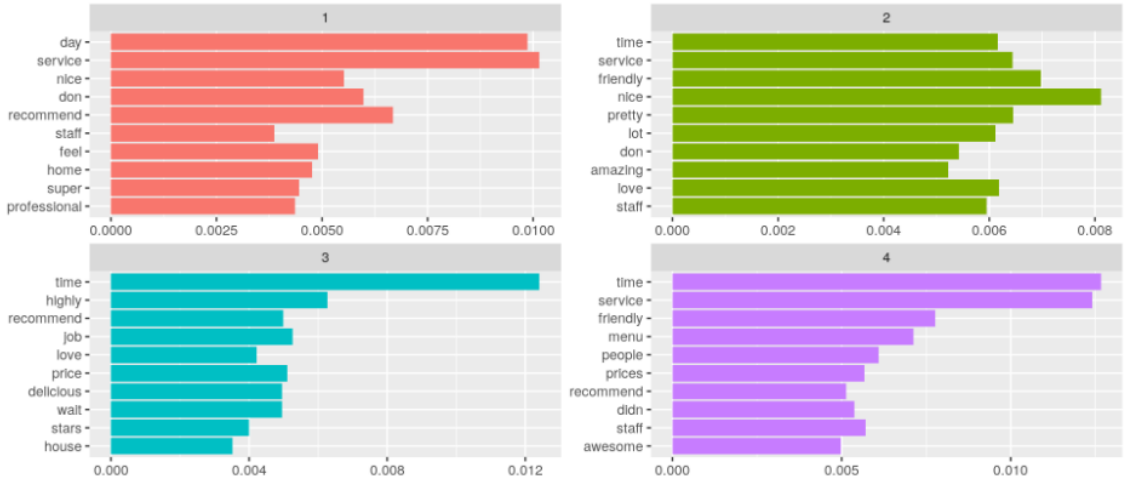
Category Key word network

Positive/Negative word emotional analysis?

Necessary cross-off!



Topic Modelling



Yelp Review Analysis

Thanks!