# Yelp Review Analysis

**Exploring**

**Single Dashboard**

**Industry Dashboard**

**City: Toronto** **Category: Food** **Sub Category: Chinese food**

# Yelp Review Analysis

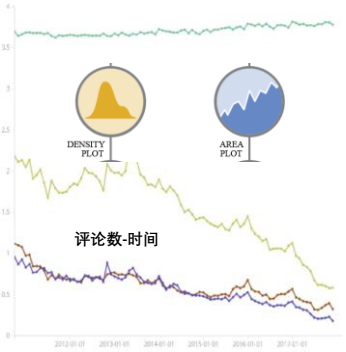
## Exploring

**Single Dashboard**

**Industry Dashboard**

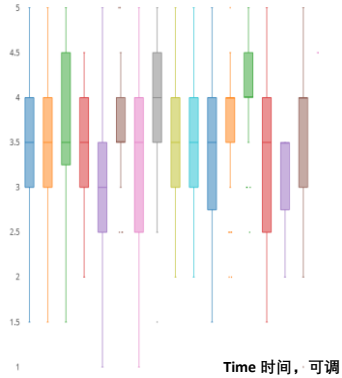
**Name: Ma's Noodle** **City: Toronto** **Category: Food** **Key: Fast food, Chinese food**

**Review number**




Time 时间，可调

**Ratings** **显示点的boxplot**



Time 时间，可调

左侧点选的时段中最高useful的rew

**PREVIEW**

在左边点选的在这里高亮 用不同颜色高亮 正面/负面词语

**Highest voted Negative**

用不同颜色高亮 正面/负面词语

左侧点选的时段中最高useful的 投票低于平均分的评论

**Highest voted Positive**

用不同颜色高亮 正面/负面词语

左侧点选的时段中最高useful的 投票高于平均分的评论

ThemeRiver/StreamGraph

如果是tips的话：
点选keywords之后在右侧显示前3个有相关关键词的时间最近的tips,在评论中将keywords标红显示，将其所在句子高亮

**Review Top 10 keywords**

1.Customer service
2.Happy hour
3.Delicious food

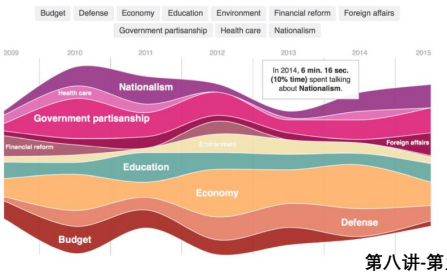
点选keywords之后在右侧显示前3个有相关关键词的useful最多的评论,在评论中将keywords标红显示，将其所在句子高亮

**Most voted related:**

review1 100 useful

review1 90 useful

review1



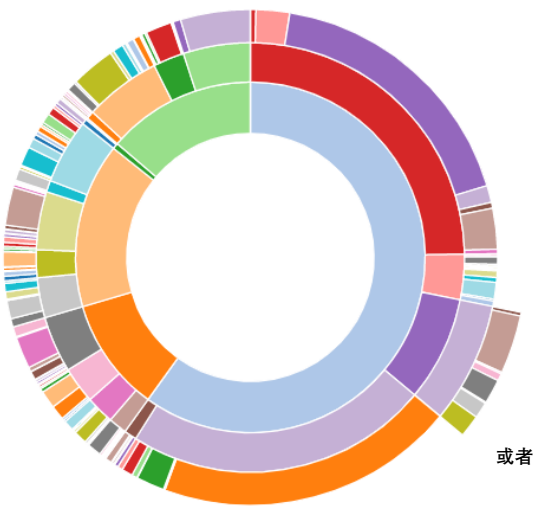

点选时段 在左侧显示详细内容

top关键词提取 选成2个单词长度！

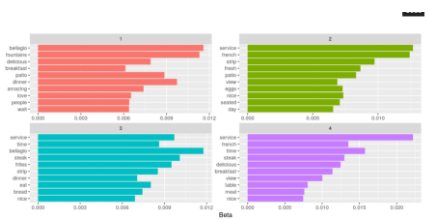
第八讲-第九讲
ppt上还有很多类似的实现，选了一些，见附录
做不出来就做带离散时间轴的词云

**Topic Key words**

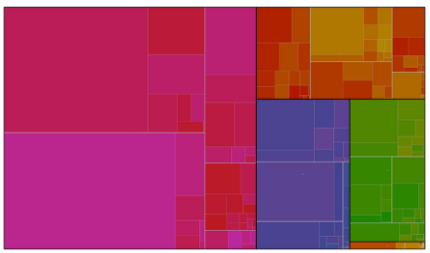
**Tips word sunburst(separated by topics)**



或者



**Tips word tree-map(separated by topics)**



暂定在单独这家餐馆的review上做，tips也可以试试

https://www.data-to-viz.com/

# Yelp Review Analysis

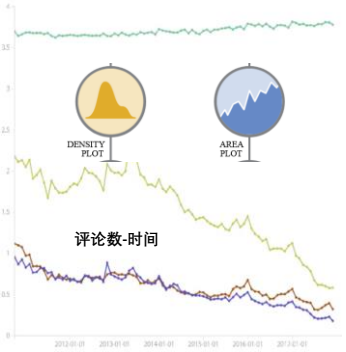
## Exploring

## Single Dashboard

## Industry Dashboard

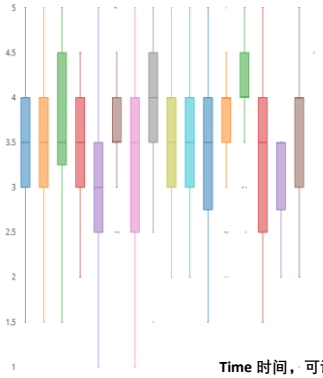
City: Toronto  Category: Food  Key: Fast food, Chinese food

Review number




Time 时间，可调

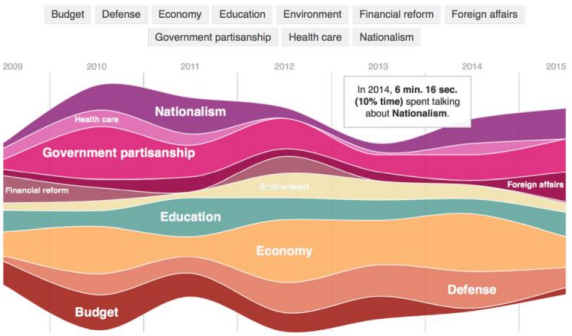
Ratings  显示点的boxplot



Time 时间，可调

本页的图和之前的x-y轴内容一致，但将对象改成同行业中同keywords(例如Chinese food)的全行业平均/top10

如果太少的话可以放到single dashboard里

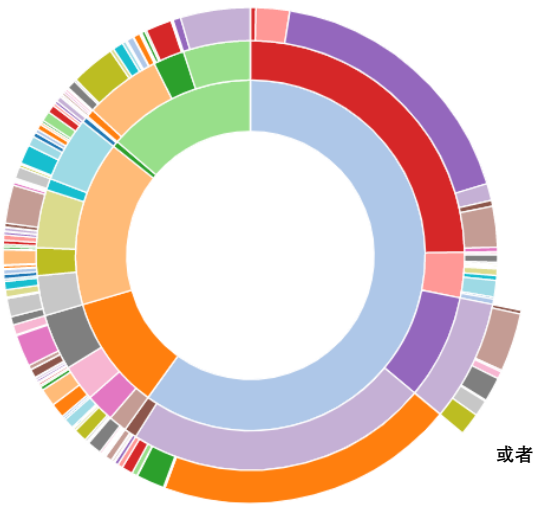



top关键词提取 选成2个单词长度！

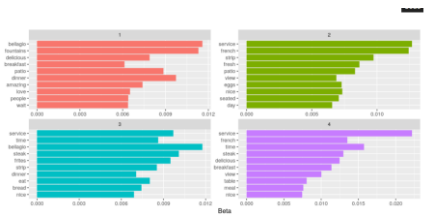
第八讲ppt上还有很多类似的实现
做不出来就做带离散时间轴的词云

Topic Key words

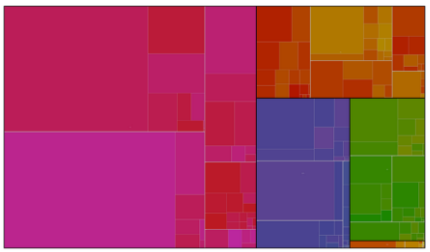
Tips word sunburst(separated by topics)



或者



Tips word tree-map(separated by topics)



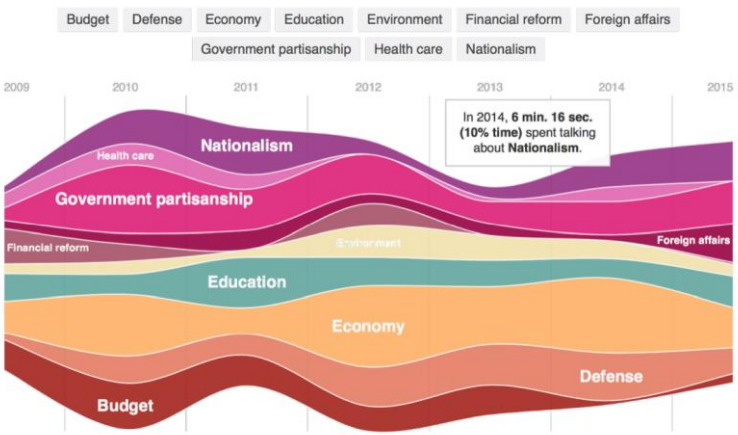
review上做，tips也可以试试

https://www.data-to-viz.com/

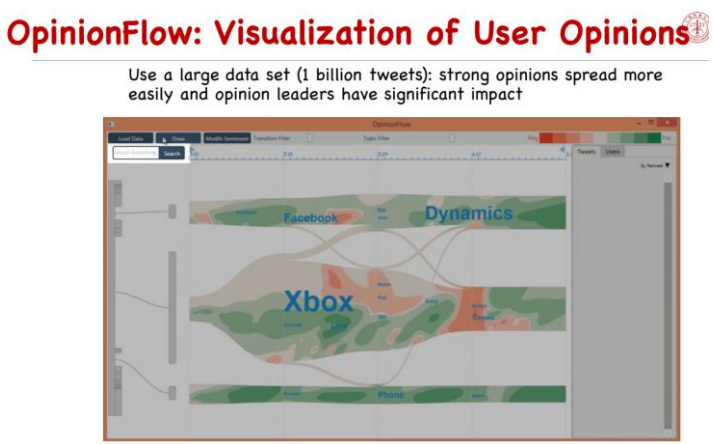
Kaggle 上的 Yelp数据集 有一些相关的工作，和问题的讨论
https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?resource=download
各种图表的实现
https://www.data-to-viz.com/

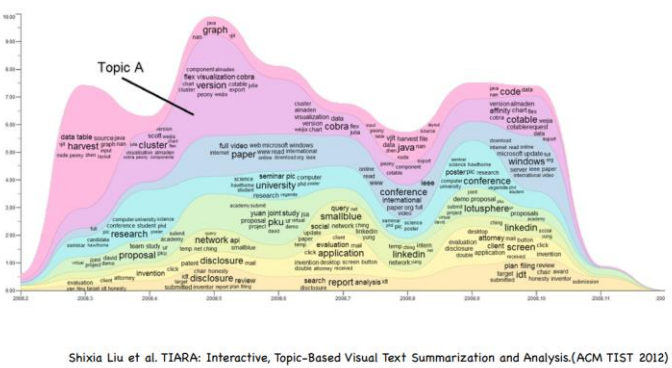
如何处理纯文本的tips、review，找出关键词频率的实现
https://www.kaggle.com/code/dehaozhang/yelp-eda/notebook
找出关键词频率的实现
https://www.kaggle.com/code/niyamatalmass/finding-the-perfect-restaurants-on-yelp
地图、处理评论
https://www.kaggle.com/code/jagangupta/what-s-in-a-review-yelp-ratings-eda#5.-Reviews-Deep-dive:
界面的布置可以参考
https://github.com/alexilyin1/yelp_shiny

限定某一个城市内的地图处理、如何找出top10商家、如何找到负面/正面词语
https://www.kaggle.com/code/ambarish/a-very-extensive-data-analysis-of-yelp




比较好的
ThemeRiver/StreamGraph




Shixia Liu et al. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis.(ACM TIST 2012)

## Dataset components

**business.json**

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

Blue: Identifying
Orange: Information
Purple: Filter

```
{
    // 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 a

    // 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"
```

# 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？

我们选择以**个体企业**为主要服务对象,帮助他们通过yelp的数据集分析自己的经营情况和同行业竞争状况

**1.以其评分、收到的建议和评价制作dashboard，**
**2.其同行业状况：整体收到的评价、建议、同类前10的状况，制作dashboard**
**3.同类店家的地理分布预览**