CSE 258

Web Mining and Recommender Systems

Assignment 2

Assignment 2

- Open-ended
- Due **Dec 4**
- Submissions should be made via gradescope

Assignment 2

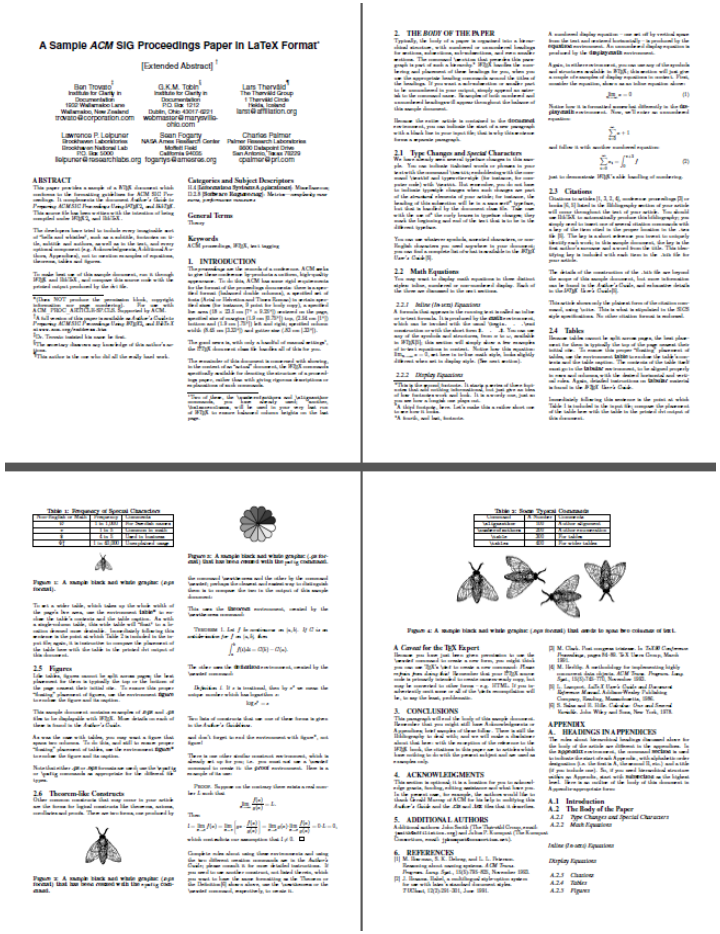
Basic tasks:

1. Identify a dataset to study and describe its basic properties
2. Identify a predictive task on this dataset and describe the features that will be relevant to it
3. Describe what model/s you will use to solve this task
4. Describe literature & research relevant to the dataset and task
5. Describe and analyze results

Assignment 2

Evaluation

- E.g. about this much:



(acm proceedings format)

https://www.acm.org/sigs/publications/proceedings-templates

Assignment 2

Teams of one to four

Assignment 2

1. Identify a dataset to study

- Beer data

(http://snap.stanford.edu/data/Ratebeer.txt.gz
http://snap.stanford.edu/data/Beeradvocate.txt.gz)

- Wine data

(http://snap.stanford.edu/data/cellartracker.txt.gz)

- Sensor data

(https://github.com/rpasricha/MetroInsightDataset)

Assignment 2

1. Identify a dataset to study

- Reddit submissions

(http://snap.stanford.edu/data/web-Reddit.html)

- Facebook/twitter/Google+ communities

(http://snap.stanford.edu/data/egonets-Facebook.html
http://snap.stanford.edu/data/egonets-Gplus.html
http://snap.stanford.edu/data/egonets-Twitter.html)

- Many many more from other sources, e.g.

http://snap.stanford.edu/data/

Use whatever you like, as long as it's **big** (e.g. 50,000 datapoints minimum)

Assignment 2

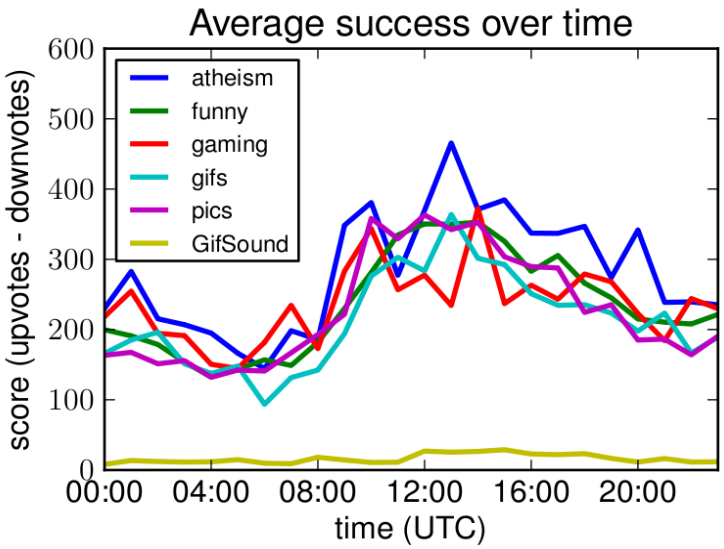
1b: Perform an **exploratory analysis** on this dataset to identify interesting phenomena

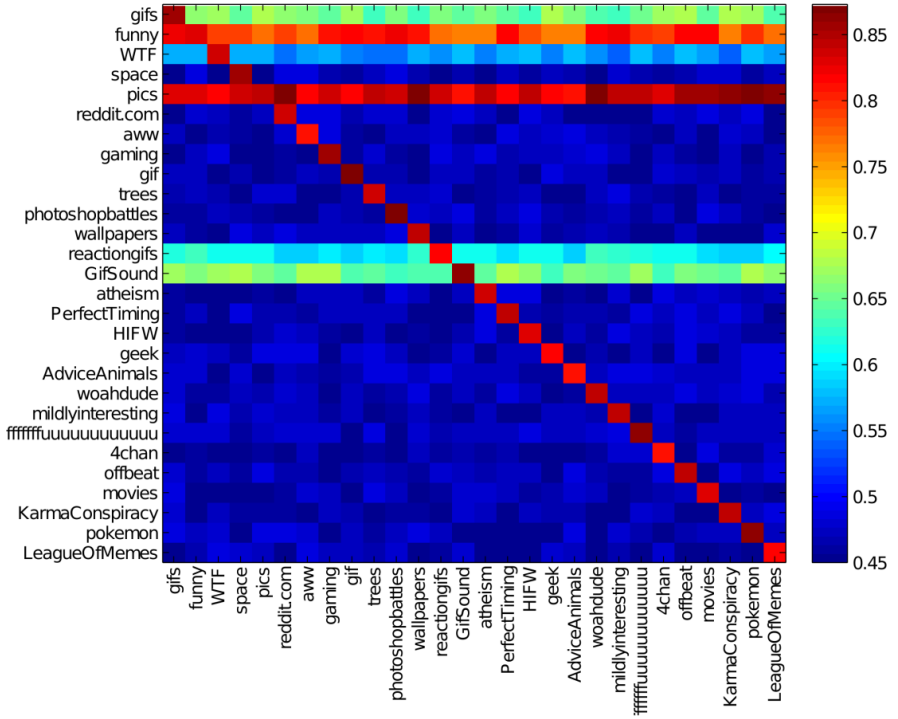
- Start with basic results, e.g. for a recommender systems type task, how many users/items/entries are there, what is the overall distribution of ratings, what time period does the dataset cover etc.

Assignment 2

1b: Perform an **exploratory analysis** of this dataset to identify interesting phenomena

e.g.





Assignment 2

2. Identify a **predictive task** on this dataset

- How will you assess the validity of your predictions and confirm that they are significant?
- Did you have to do pre-processing of your data in order to obtain useful features?
- How do the results of your exploratory analysis justify the features you have chosen?

Assignment 2

3. Select/design an appropriate model

- How will you evaluate the model? Which models from class are relevant to your predictive task, and why are other models inappropriate?
- It's totally fine here to implement a model that we covered in class, e.g. for a classification task you could implement svms+logistic regression+naïve Bayes
- You should also compare the results of different feature representations to identify which ones are effective
- What are the relevant baselines that can be compared?
- If you used a complex model, how did you optimize it?
  - What issues did you face scaling it up to the required size?
  - Any issues overfitting?
  - Any issues due to noise/missing data etc.?

Assignment 2

4. Describe related literature

- If you used an existing dataset, where did it come from and how was it used there?
- What other similar datasets have been used in the past and how?
- What are the state-of-the-art methods for the prediction task you are considering? Were you able to borrow any ideas from these works for your model? What features did they use and are you able to use the same ones?
- What were the main conclusions from the literature and how do they differ from/compare to your own findings?

Assignment 2

5. Describe your results

- Of the different models you considered, which of them worked and which of them did not?
- What is the interpretation of the parameters in your model? Which features ended up being predictive? Can you draw any interesting conclusions from the fitted parameters?

Assignment 2

Example

Maybe I want to use **restaurant data** to build a model of people's tastes in different locations

Assignment 2

1. Perform an **exploratory analysis** of this dataset to identify interesting phenomena

- How many users/items/ratings are there? Which are the most/least popular items and categories?
- What is the geographical spread of users, items, and ratings?
- Do people give higher/lower ratings to more expensive items, or items in certain countries/locations?

Assignment 2

2. Identify a **predictive task** on this dataset

- Predict what rating a person will give to a business based on the time of year, the past ratings of the user, and the geographical coordinates of the business
- Predict which businesses will succeed or fail based on its geographical location, or based on its early reviews
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?

Assignment 2

2b. Identify features that will be relevant to the task at hand

- Ratings, users, geolocations, time
- Ratings as a function of price
- Ratings as a function of location
  - How to represent location in a model? Just using a linear predictor of latitude/longitude isn't going to work…

Assignment 2

3. Select an appropriate model

- Some kind of latent-factor model
- How to incorporate the geographical term? Should we cluster locations? Use the location as a regularizer? (etc.)
- How can we optimize this (presumably complicated) model?

Assignment 2

4. Describe related literature

- Relevant literature or predicting ratings
- Literature on using geographical features for various predictive tasks
- Literature on predicting long-term outcomes from time series data
- Literature on predicting future ratings from early reviews, herding etc.

Assignment 2

5. Describe results and conclusions

- Did features based on geographical information help? If not why not?
- Which locations are the most price sensitive according to your predictor?
- Do people prefer restaurants that are unlike anything in their area, or restaurants which are exactly the same as others in their area?

Assignment 2

Example 2

Maybe I want to use **reddit data** to see what makes submissions successful

(http://snap.stanford.edu/data/web-Reddit.html)

Assignment 2

1. Perform an **exploratory analysis** of this dataset to identify interesting phenomena

- How many users/submissions are there? How does activity differ across subreddits?
- What times of day are submissions most commented on or most rated?
- Do people give more/fewer votes to submissions that have long/short titles, or which use certain words?

Assignment 2

2. Identify a **predictive task** on this dataset

- Predict whether a post will have a large number of comments or a high rating
- Predict whether there will be a large *discrepancy* between the number of comments and the positivity of ratings a post receives
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?

Assignment 2

2b. Identify features that will be relevant to the task at hand

- Votes, users, subreddits, time
- Resubmissions of the same content & the success or failure of previous submissions
- Text of the post title

Assignment 2

3. Select an appropriate model

- Some kind of regression
- Need to use gradient descent or is there a closed-form solution?
- What are the hyperparameters and how do we regularize?
- How can you incorporate the temporal terms?

Assignment 2

4. Describe related literature

- Relevant literature or predicting votes on Reddit
- Literature on virality in social media
- Literature on using text for predictive tasks
- Literature on temporal forecasting or user preference modeling

Assignment 2

5. Describe results and conclusions

- What features helped you to predict whether content would be controversial or not?
- Does the text of the title help to predict whether a submission will be controversial or get many comments but a low vote?
- Which subreddits generate more controversial content than others?

Assignment 2

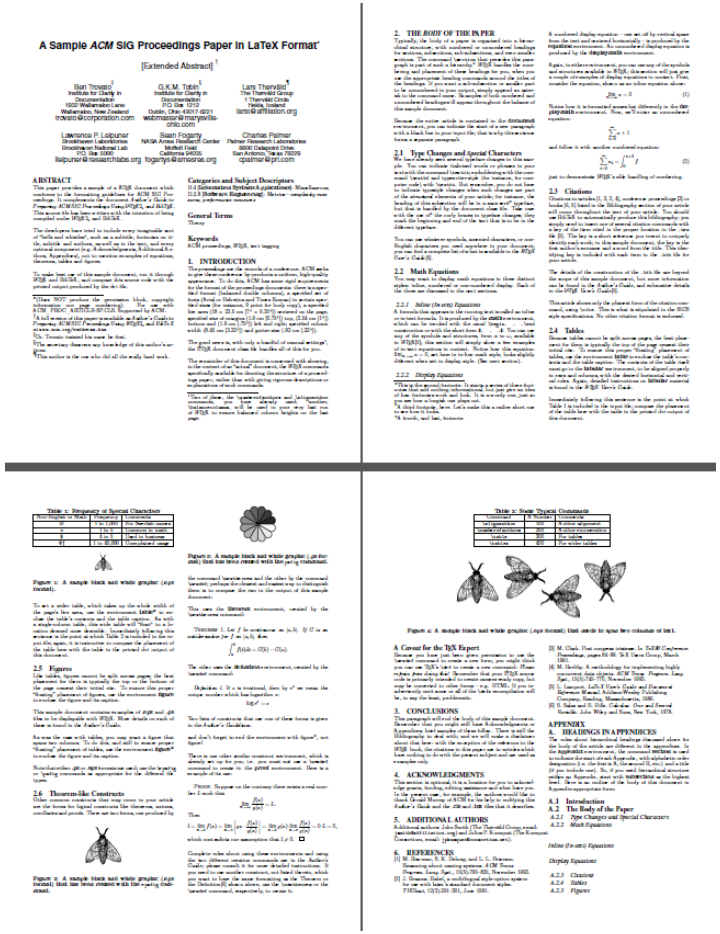
Evaluation

- These 5 sections will be worth (roughly) 5 marks each (for a total of 25% of your grade)
- Assignments can be done **in groups of up to 3 (or 4).** The marking scheme is the same regardless of group size.
- Length is not strict, but should be about 4 pages in small-font double-column format.

Assignment 2

Evaluation

- E.g. about this much:



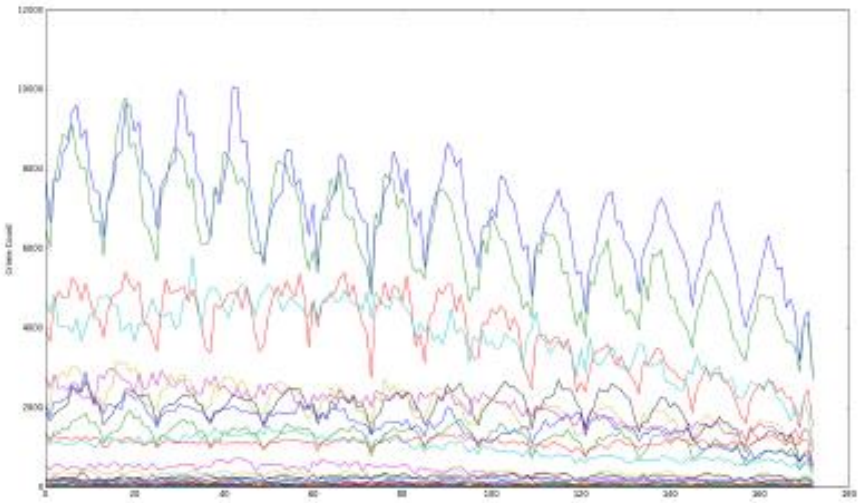
(acm proceedings format)

https://www.acm.org/sigs/publications/proceedings-templates

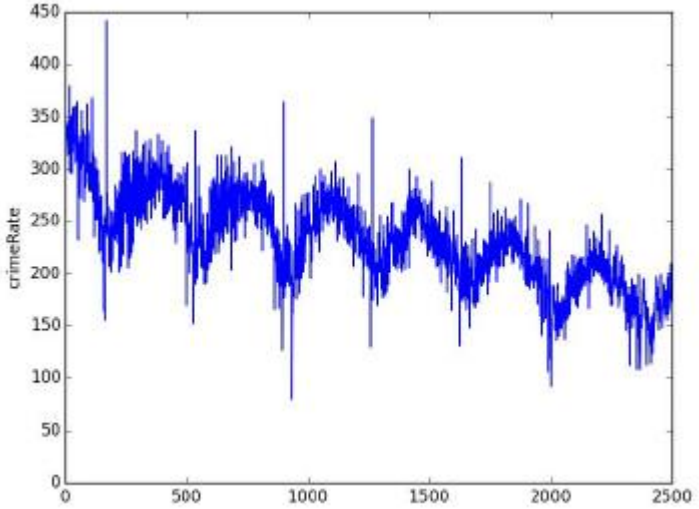
Data Mining and Predictive Analytics

Assignment 2 – examples of previous assignments

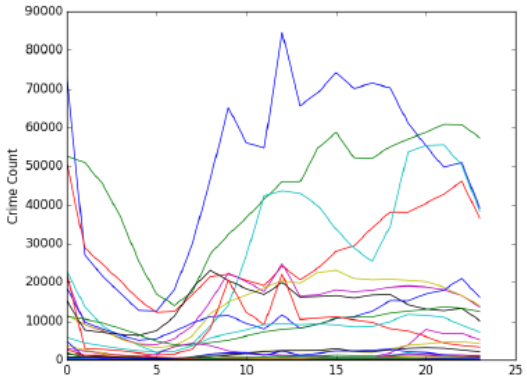
Crime (Chicago)



Over 15 years



Over 7 years

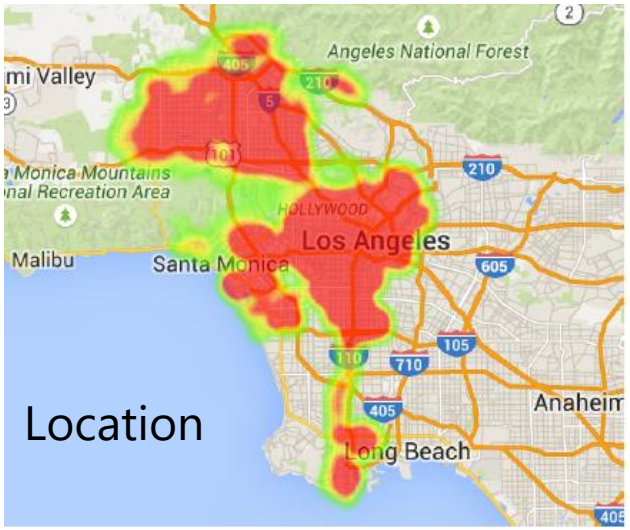


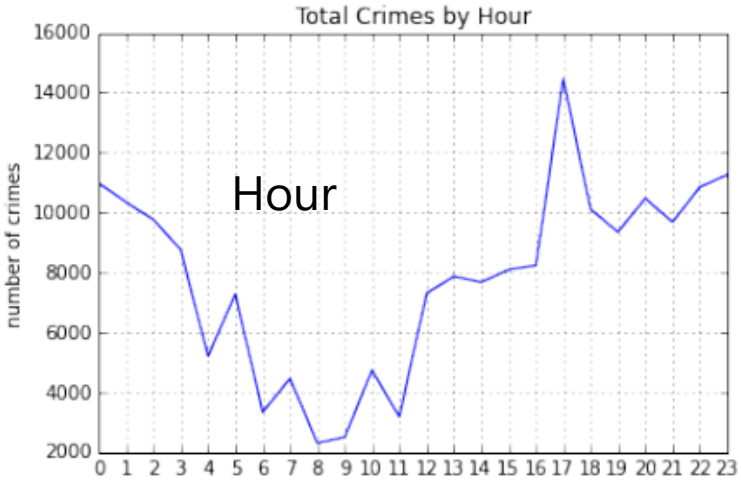
Hour of the day

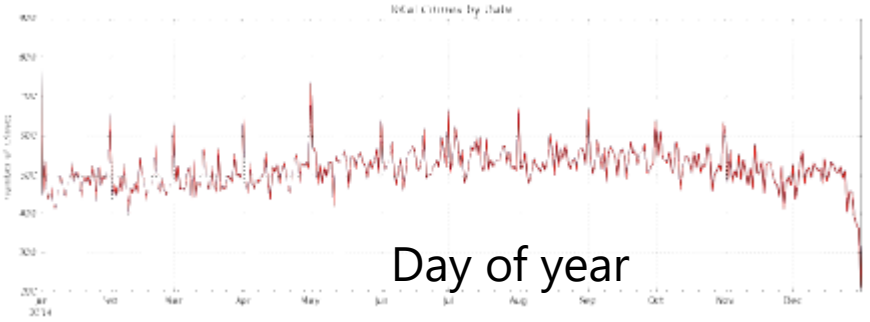
Goal: to predict the number of incidents of crime on a given day

Joshua Wheeler, Nathan Moreno, Anjali Kanak

Crime (Los Angeles)

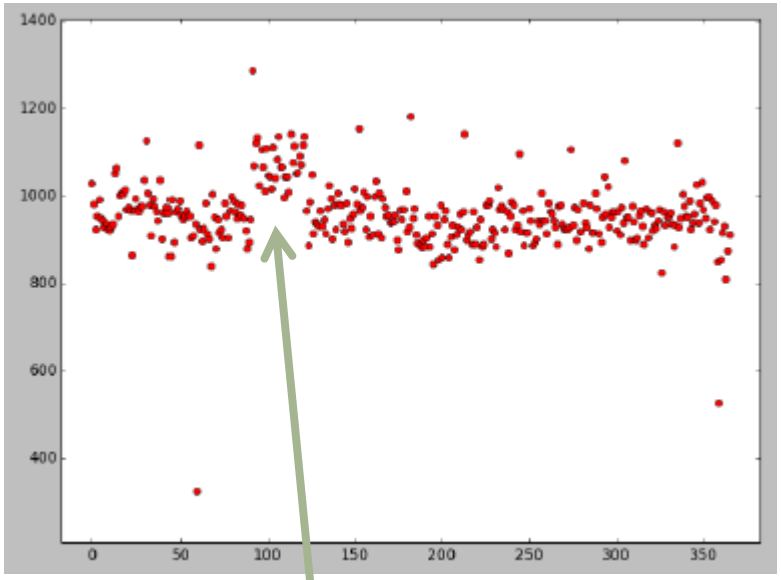




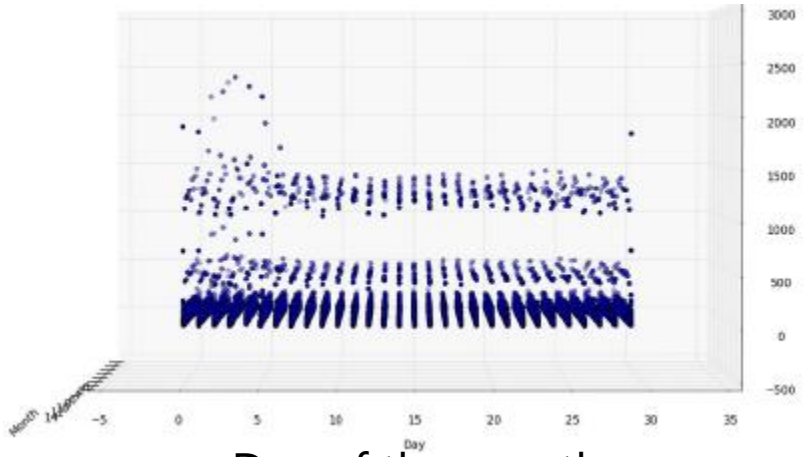


Hye Lim Park

Crime (San Diego)

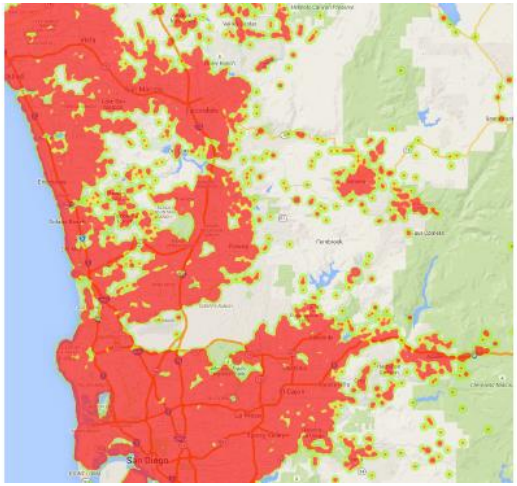


April 10th



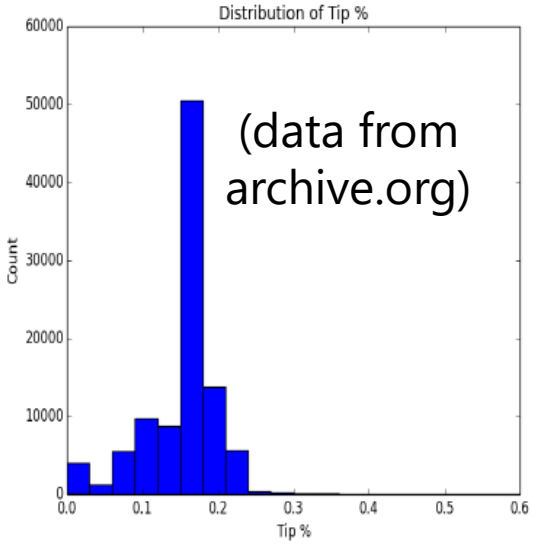
Day of the month

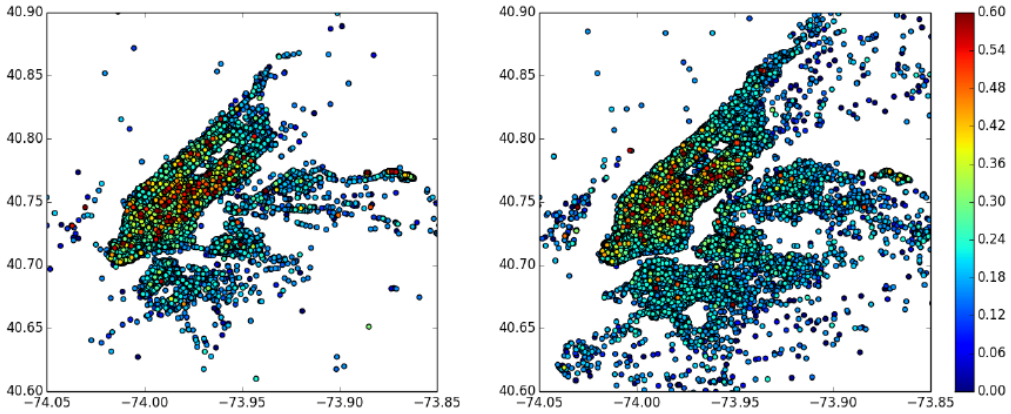
Location:



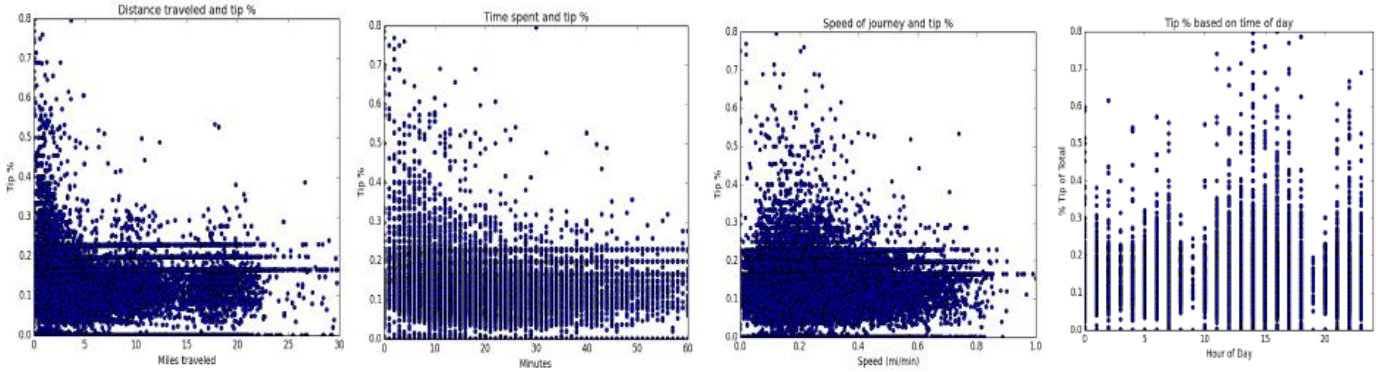
Steve Morlan, David Hart, Ilia Shumailov

Predicting Taxi Tip-Rates in NYC





(pickup and dropoff)



Distance, time taken, speed, and time of day (also on geo)

Sahil Jain, Alvin See, Anish Shandilya

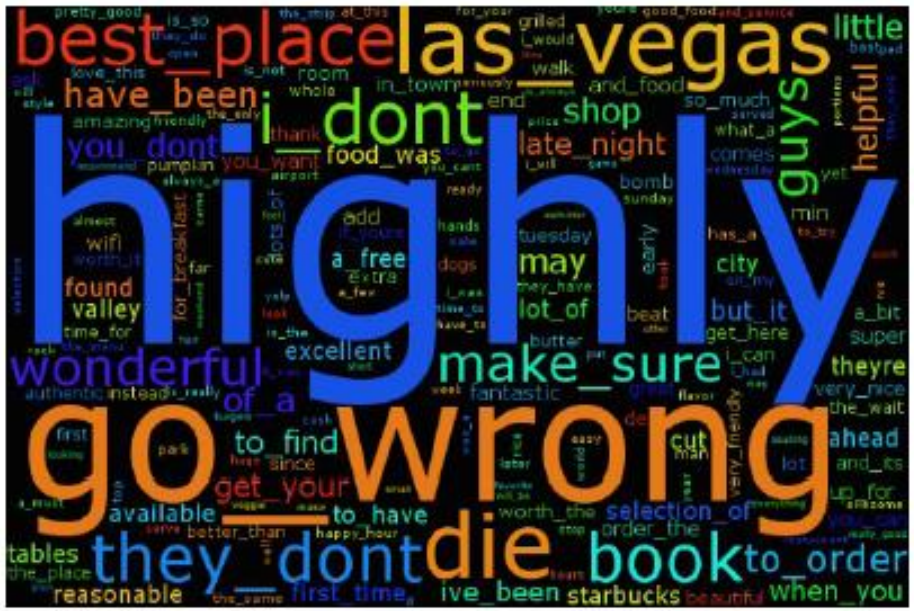
Wordles!

Amazon Gourmet foods:
Michael Tran

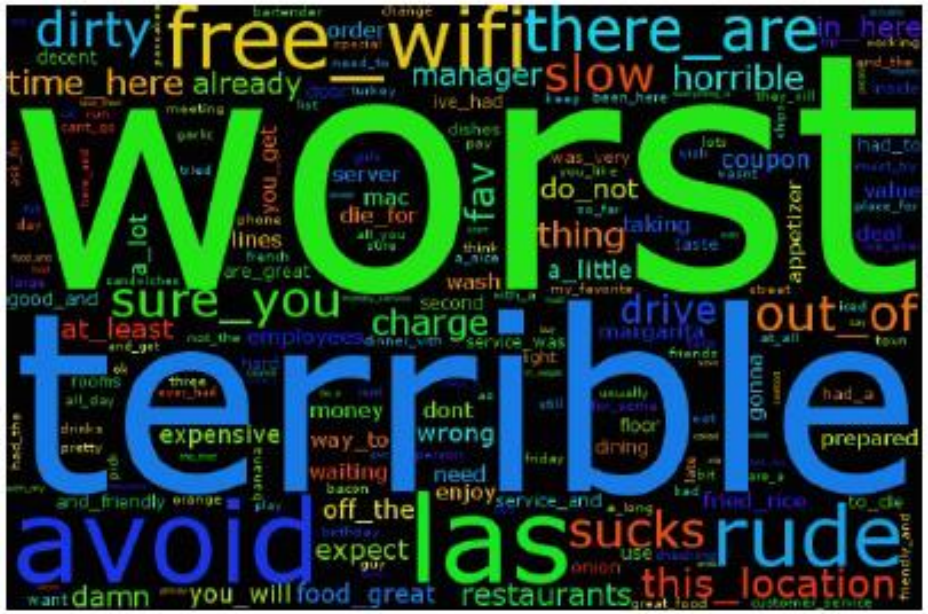
Wordles!

Amazon Clothing:
Hen Su Choi Ortiz, Rajat Shah

Wordles!



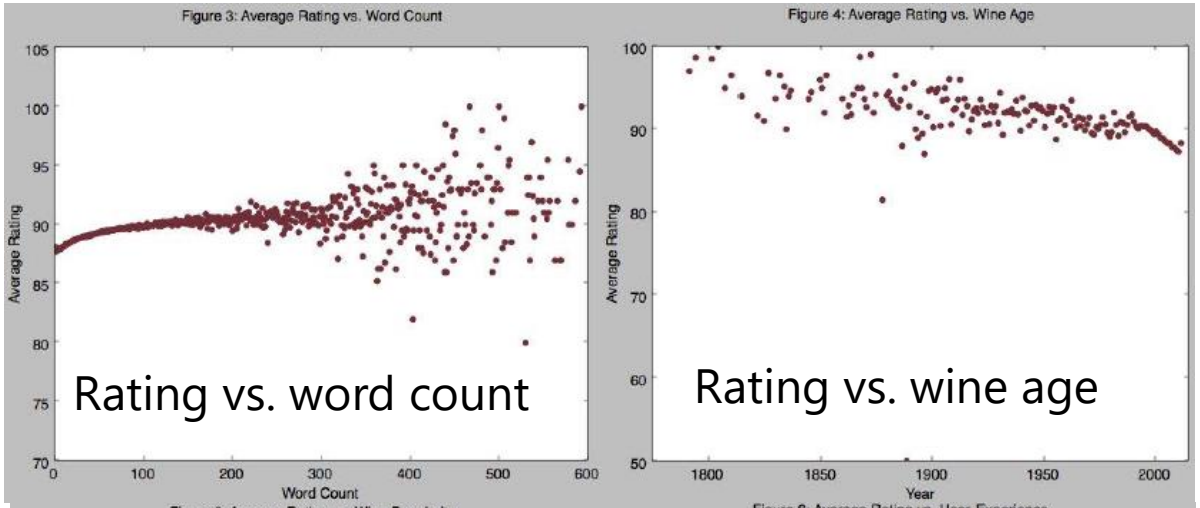
Positive

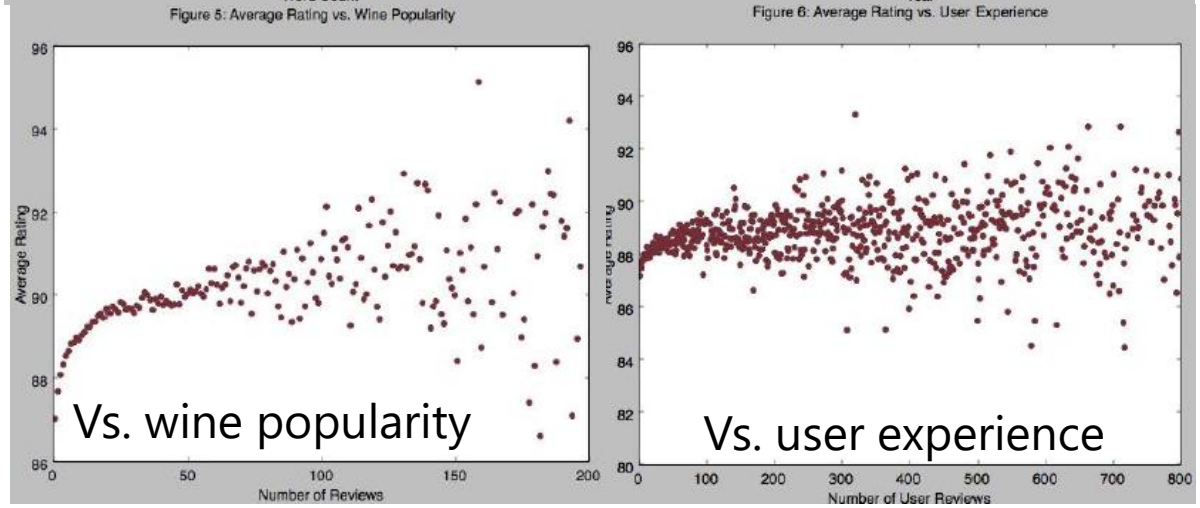


Negative

Yelp:
Angelique De Castro, Andrew Du, Aieswaryasayee Manicka

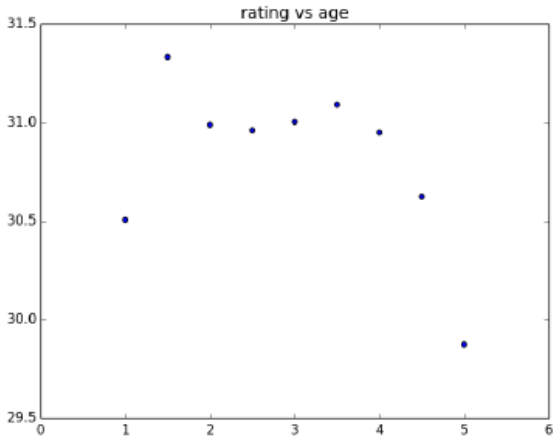
Wine ratings



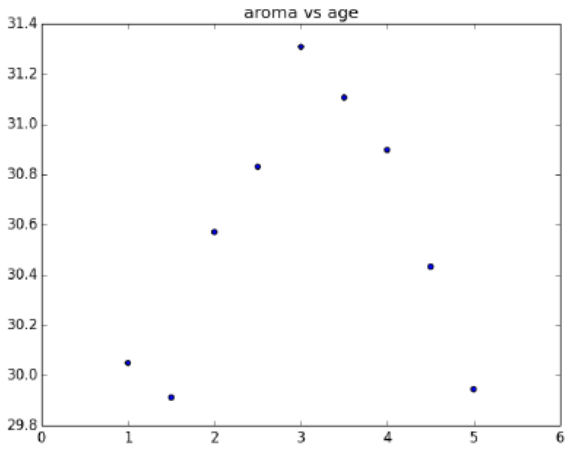


Forrest Delavega, Ryan Almodovar

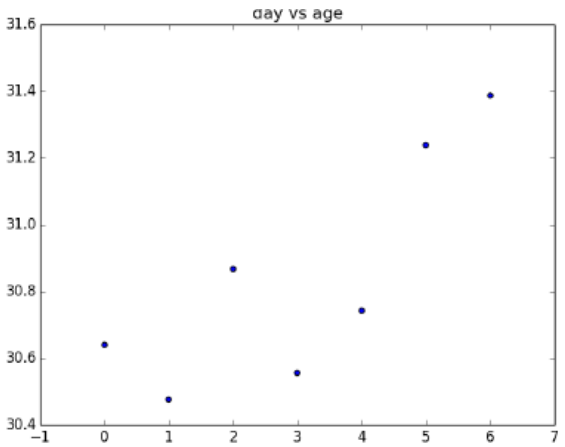
User age



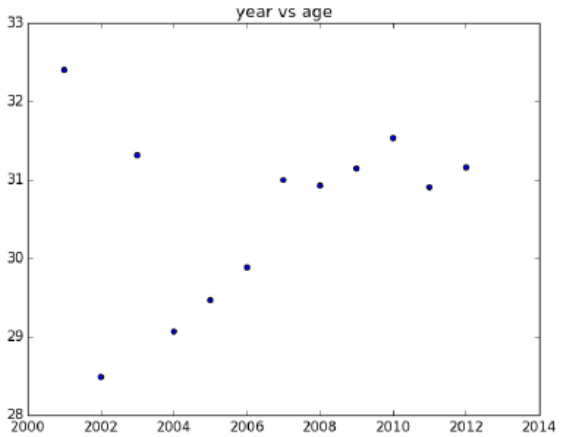
Rating vs. age



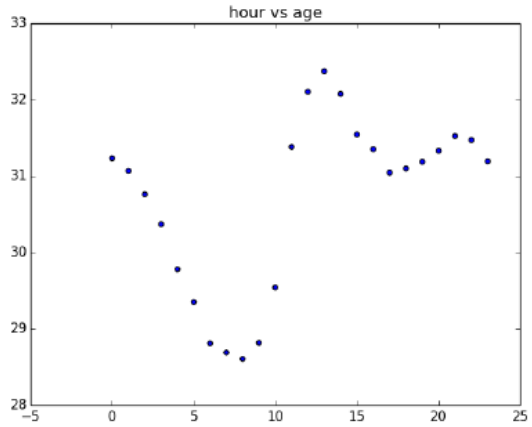
Aroma vs. age



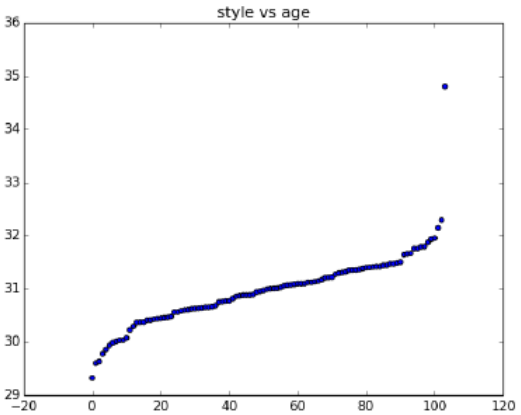
Day of week vs. age



Year vs. age



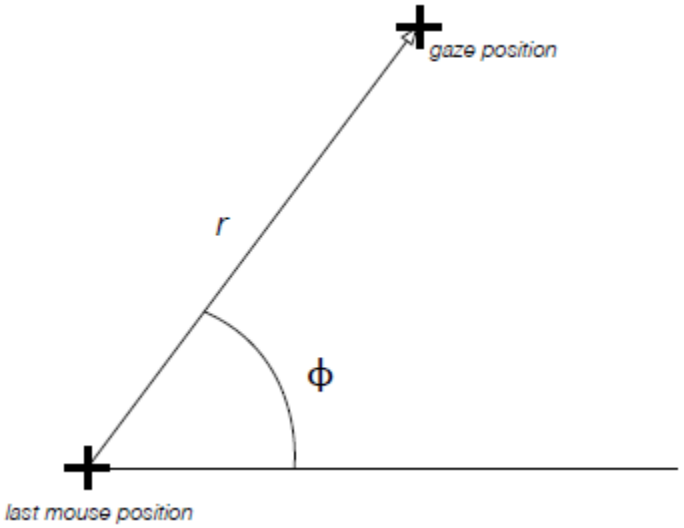
Hour of day vs. age

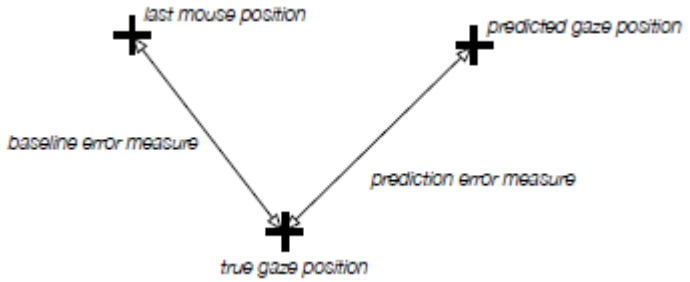


Category vs. age

Joseph Luttrell, Spenser Cornett

Gaze prediction





$B_x :=$ x-bucket membership vector

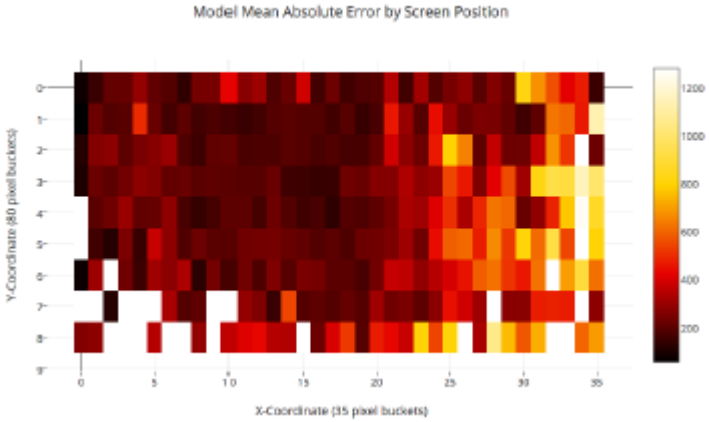
$B_y :=$ y-bucket membership vector

$\nabla M_x :=$ gradient in x direction

$\nabla M_y :=$ gradient in y direction

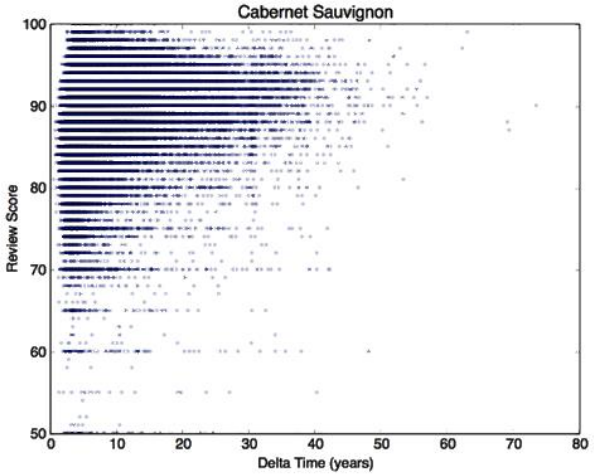
$\nabla M_x^2 :=$ 2nd gradient in x direction

$\nabla M_y^2 :=$ 2nd gradient in y direction

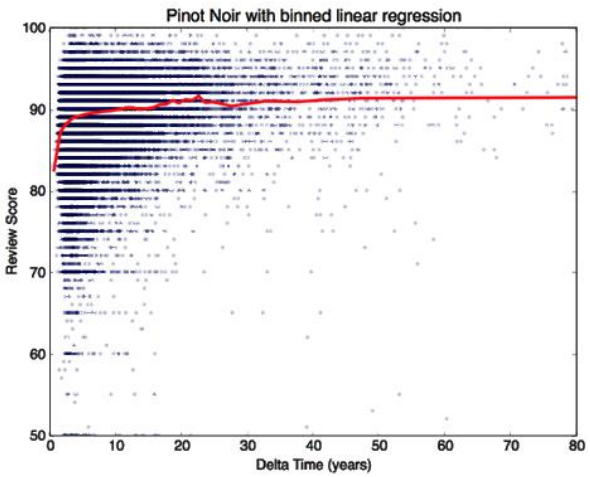


Joshua Toenyes, Samuel Vange, Saif Chaudhry

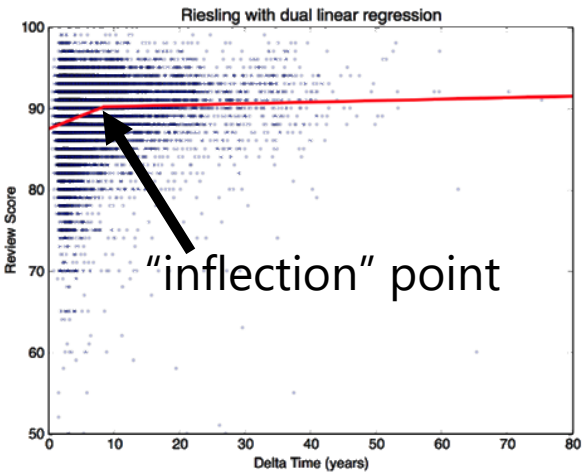
Assignment 2



Raw rating data



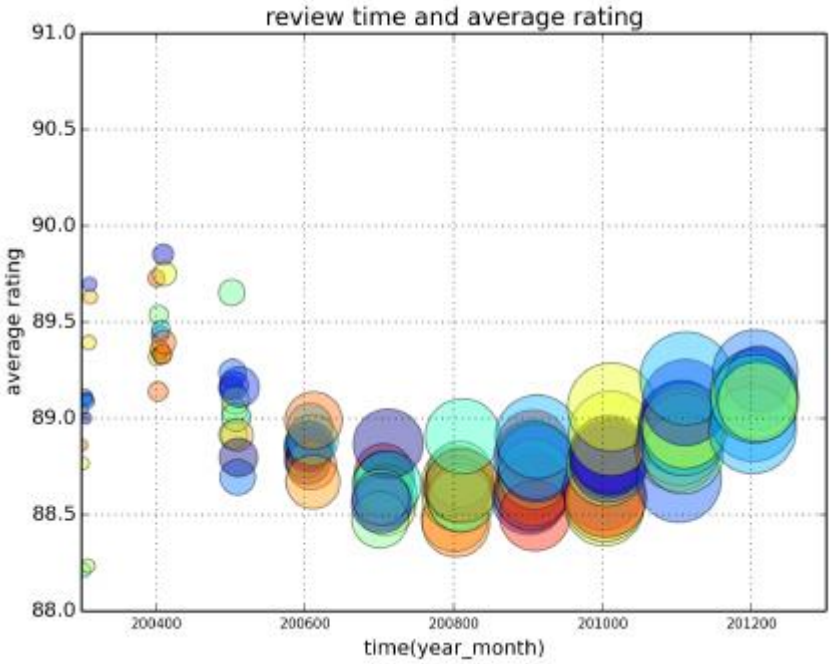
binned regression



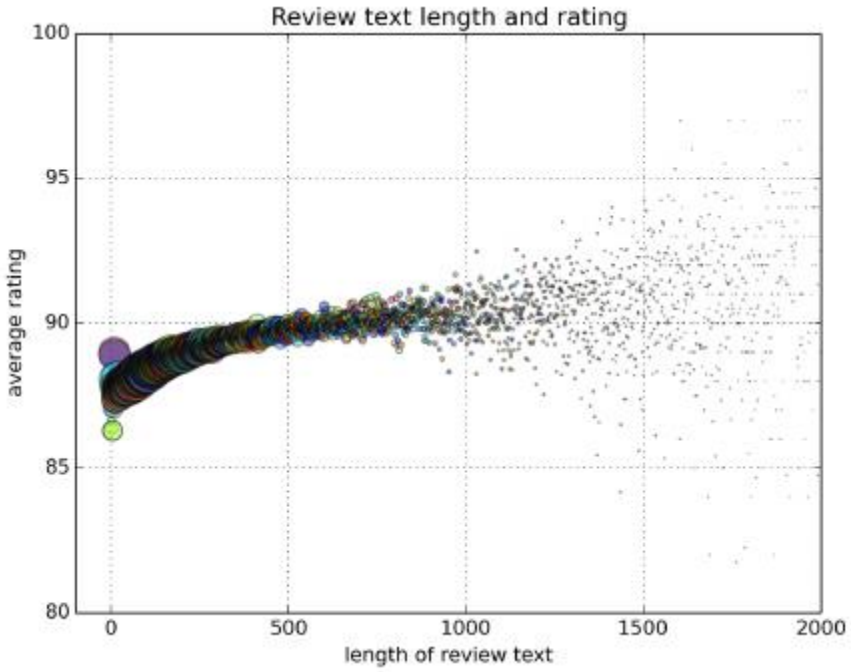
dual regression

Andrew Prudhomme – "Finding the Optimal Age of Wine"

Assignment 2



ratings vs. time



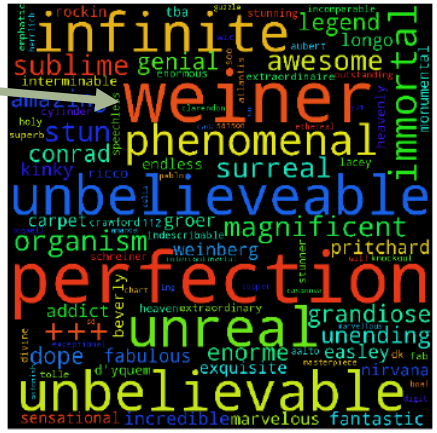
ratings vs. review length

Ruogu Liu – "Wine Recommendation for CellarTracker"

Assignment 2

?

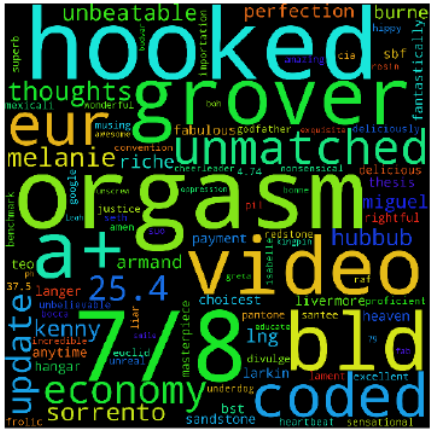
cellartracker:



positive words in wine reviews

negative words in wine reviews

RateBeer:



positive words in beer reviews

negative words in wine reviews

Ben Braun & Robert Timpe – "Text-based rating predictions from beer and wine reviews"

Assignment 2

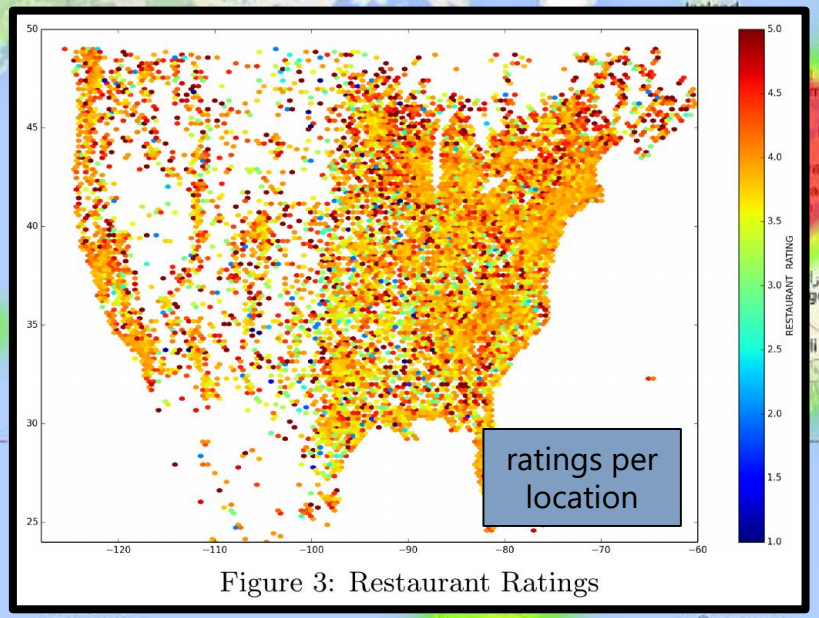


Figure 3: Restaurant Ratings

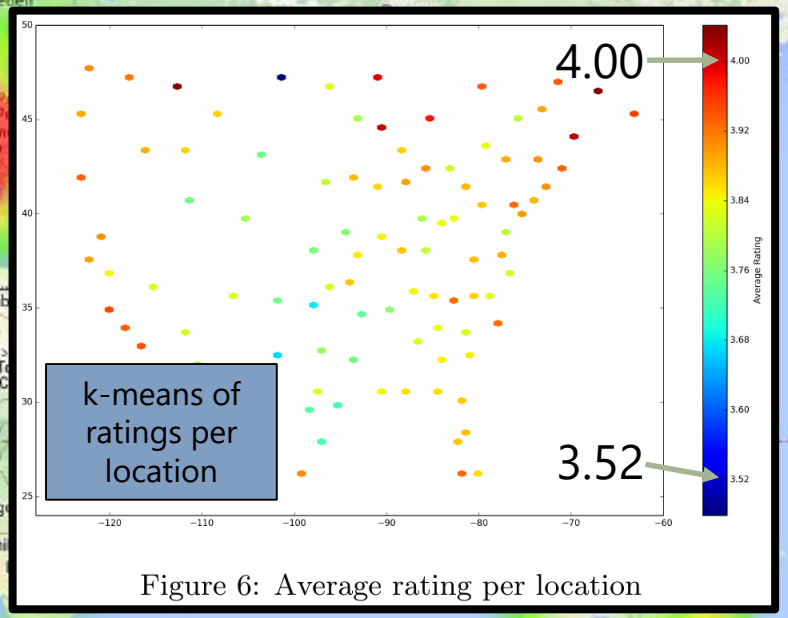
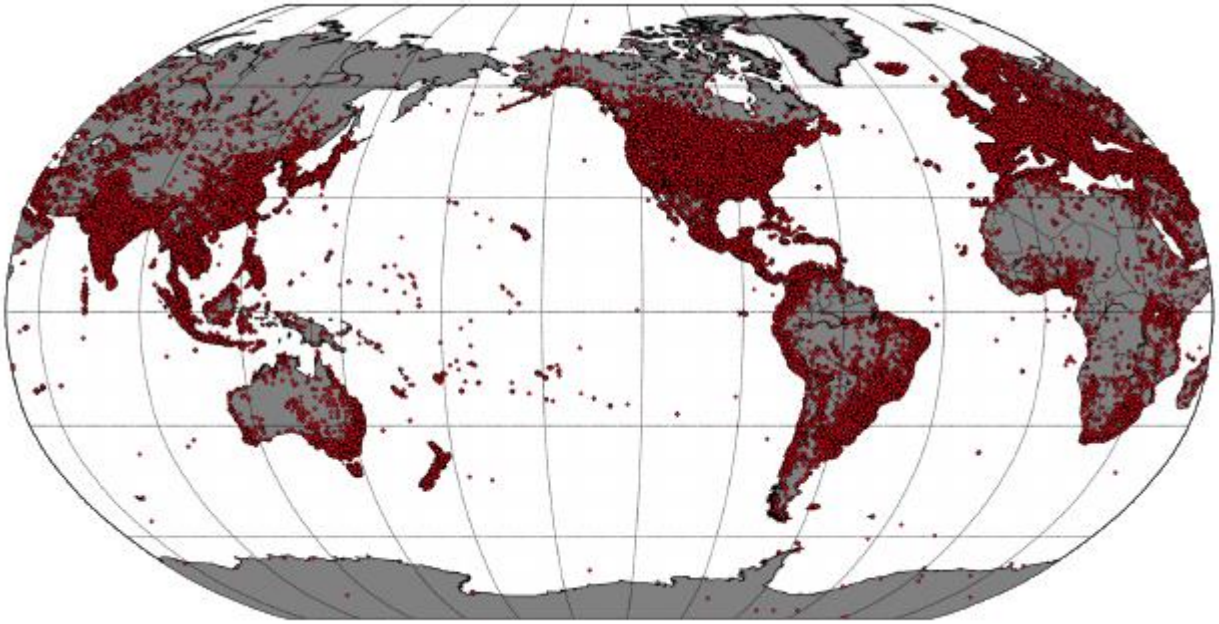


Figure 6: Average rating per location

Diego Cedillo & Idan Izhaki – "User Score for Restaurants Recommendation System"

Assignment 2



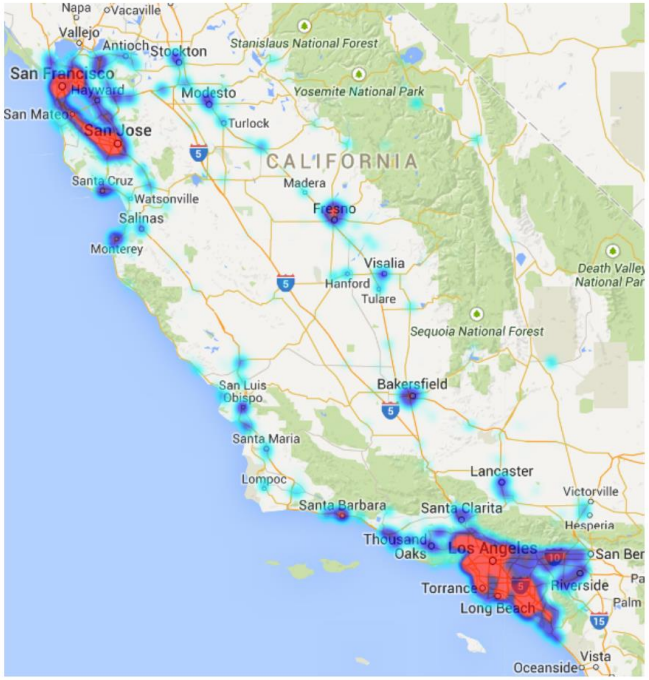
$$\widehat{r_{ui}} = \mu + b_u + b_i + (q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n|)^T p_u$$

set of geographic neighbours

impact of neighbours

Long Jin & Xinchi Gu – "Rating Prediction for Google Local Data"

Assignment 2



"Fitness"	*"Italian Restaurants"*	*"Airport & Rentals"*	*"Computer Repairs"*	*"Mexican"*
gym	food	san	computer	food
training	restaurant	francisco	store	mexican
fitness	wine	car	phone	tacos
classes	menu	airport	system	burrito
equipment	great	jose	buy	good
class	delicious	time	laptop	salsa
life	service	rental	apple	taco
great	dinner	driver	repair	chips
workout	dishes	service	problem	burritos
weight	excellent	bus	back	fish
ve	dining	shuttle	fixed	chicken
work	meal	taxi	pc	place
body	italian	trip	drive	delicious
yoga	experience	city	price	love
trainers	amazing	cab	data	fresh
people	wonderful	lax	fix	great
years	atmosphere	area	iphone	beans
feel	small	experience	screen	restaurant
instructors	decor	company	bought	asada

Topic model from Google Local business reviews

Mohit Kothari & Sandy Wiraatmadja – "Reviews and Neighbors Influence on Performance of Business"

Assignment 2

Wikispeedia navigation traces:

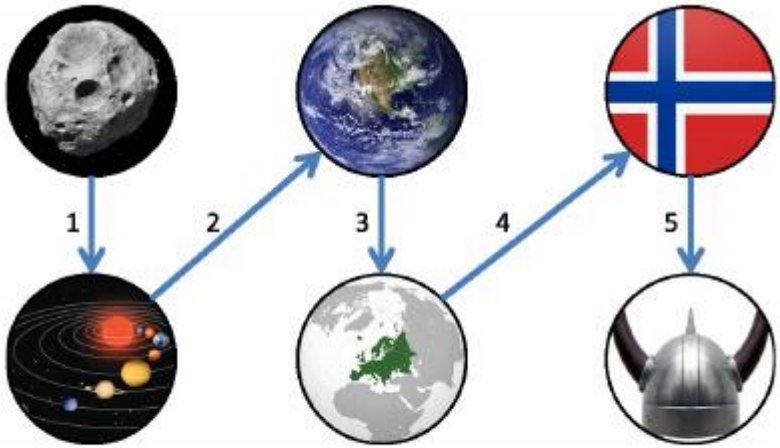


Figure 5: Graph of a complete path

	Average Click	Average Time
Finish Path	4.72	158.27
Finished Path Back	6.75	158.31
Unfinished Path	2.97	835.29
Unfinished Path Back	5.2	836.00

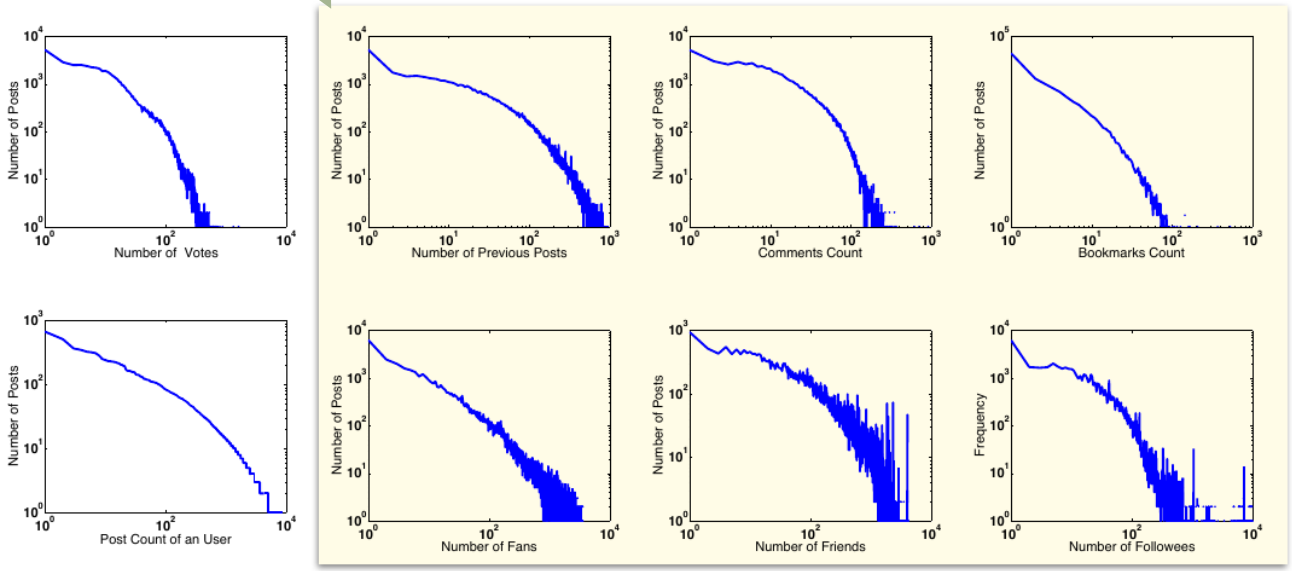
Shelby Thomas & Moein Khazraee – "Determining Topics in Link Traversals through Graph-Based Association Modeling"

Assignment 2

Images from Chictopia

Tags	electric, every day, summer, cute, T-shirt, chic
Clothes	Chartreuse Uniqlo Socks Light Blue Uniqlo T-Shirt Bubble Gum Tie-Ups Belt White Christian Louboutin Heels
User Information	1369 friends 15 followees 2245 fans
Popularity	129 votes 62 comments 15 bookmarks

Power laws!



Wei-Tang Liao & Jong-Chyi Su – "Image Popularity Prediction on Social Networks"

Questions?