

CSE 190

Data Mining and Predictive Analytics

Assignment 2

Assignment 2

- Open-ended
- Due **June 2** (four weeks from two days ago)
- Submissions should be made electronically to Long Jin (longjin@cs.ucsd.edu)

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 literature & research relevant to the dataset and task
4. Describe and analyze results

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>)

- Google Local (Maps & Restaurants)

(<http://jmcauley.ucsd.edu/data/googlelocal.tar.gz> - warning: kind of huge)

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

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

- How will you evaluate the model? Which models from class are relevant to your predictive task, and why are other models inappropriate?
- What are the relevant baselines that can be compared?
- How will you assess the validity of your predictions and confirm that they are significant?
- 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
- 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. 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

4. Describe your results

- 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.?
- 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

(<http://jmcauley.ucsd.edu/data/googlelocal.tar.gz>)

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. Describe related literature

- Relevant literature on 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

4. 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

More examples

A similar type of project from Stanford's
"Social and Information Network Analysis"
course:

[http://snap.stanford.edu/class/cs224w-
2013/projects.html](http://snap.stanford.edu/class/cs224w-2013/projects.html)

Assignment 2

More examples

Last quarter's graduate course (cse 255)

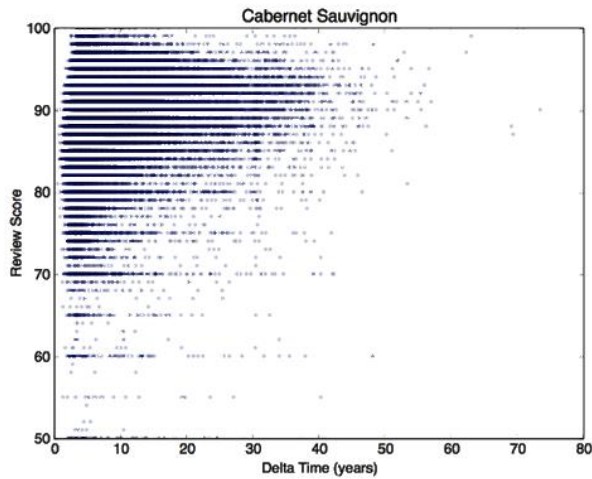
<http://cseweb.ucsd.edu/~jmcauley/cse255/projects/>

Assignment 2

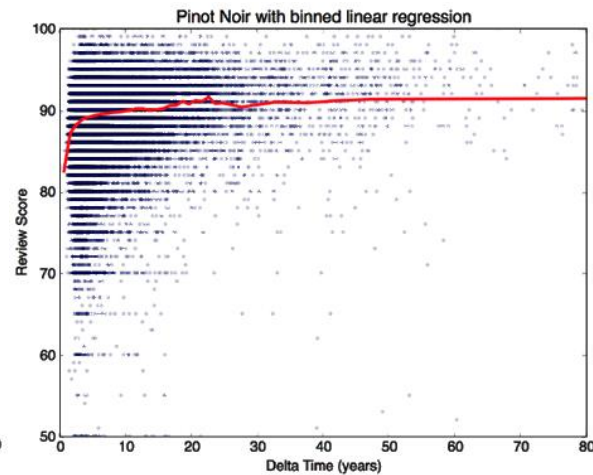
Evaluation

- These 4 sections will be worth (roughly) 5 marks each (for a total of 20% of your grade)
- Assignments can be done **in groups of up to 3**. 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.

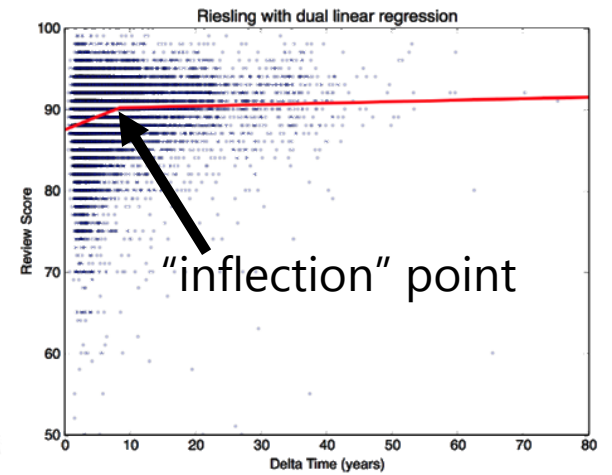
Assignment 2



Raw rating data

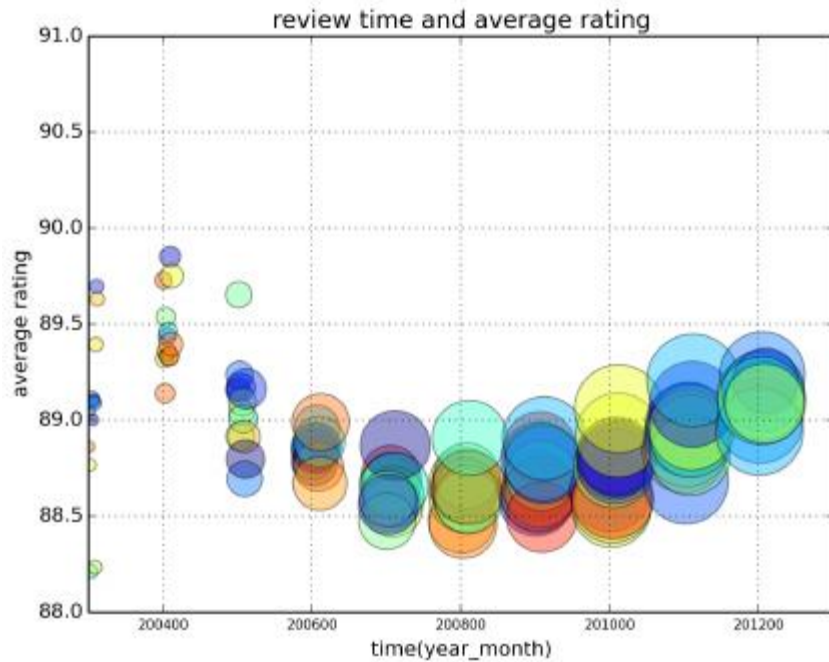


binned regression

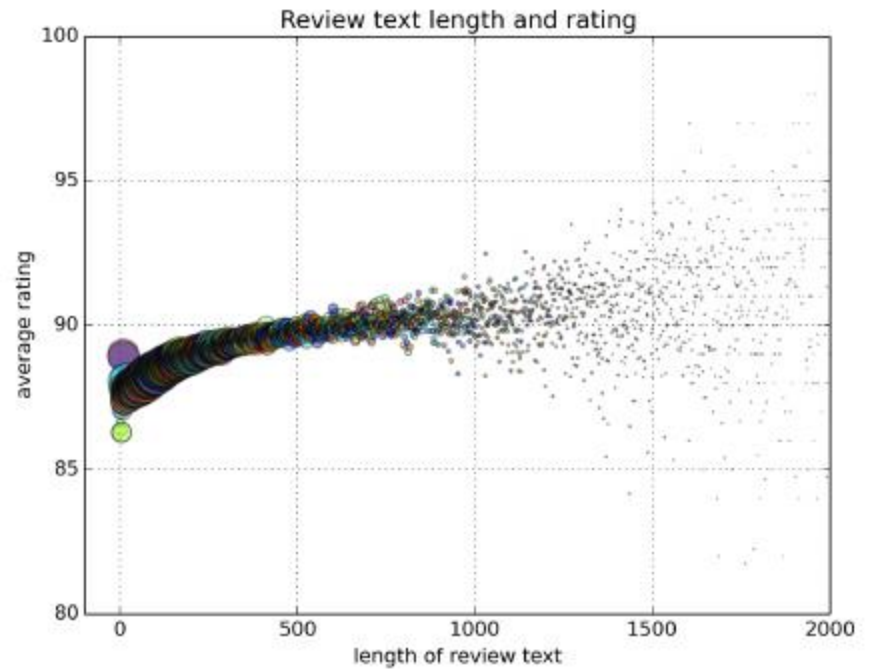


dual regression

Assignment 2



ratings vs. time

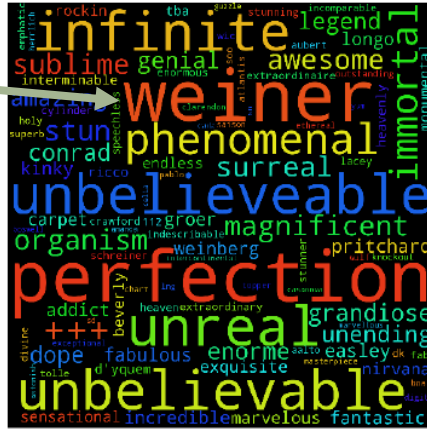


ratings vs. review length

Assignment 2

?

cellartracker:



positive words in wine reviews

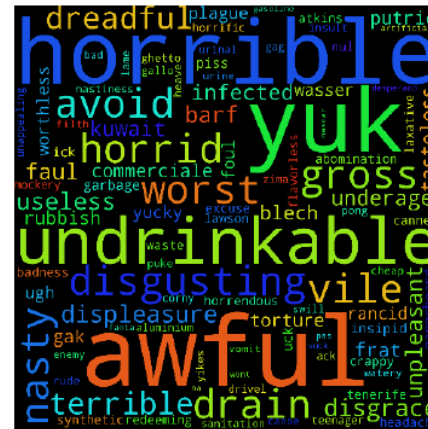


negative words in wine reviews

RateBeer:

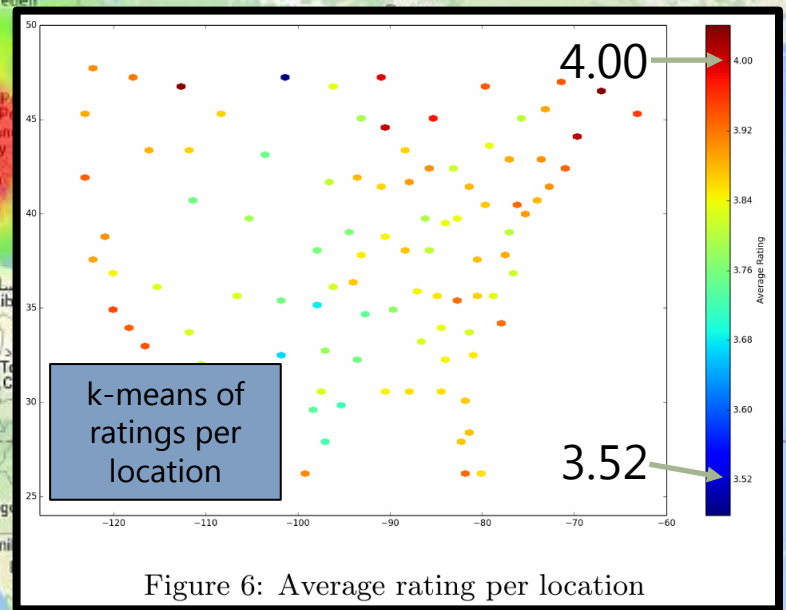
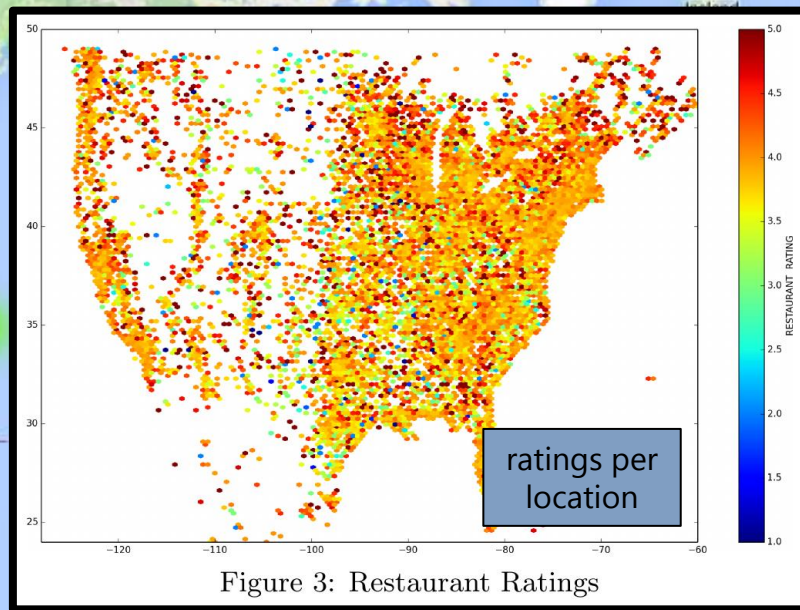


positive words in beer reviews

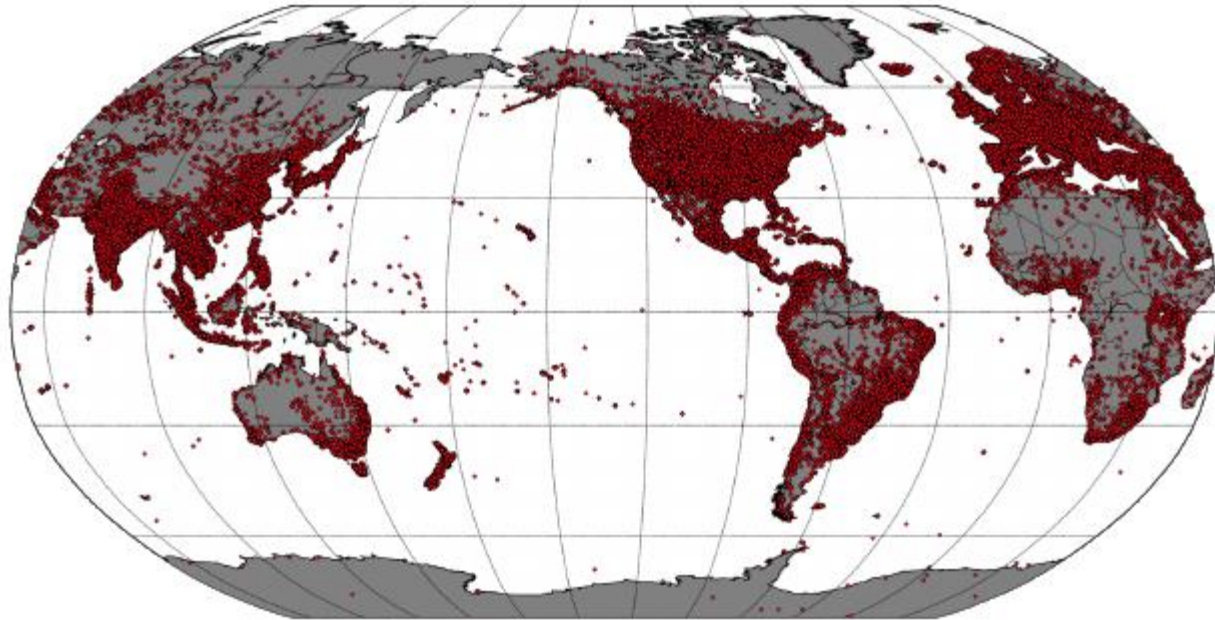


negative words in wine reviews

Assignment 2



Assignment 2

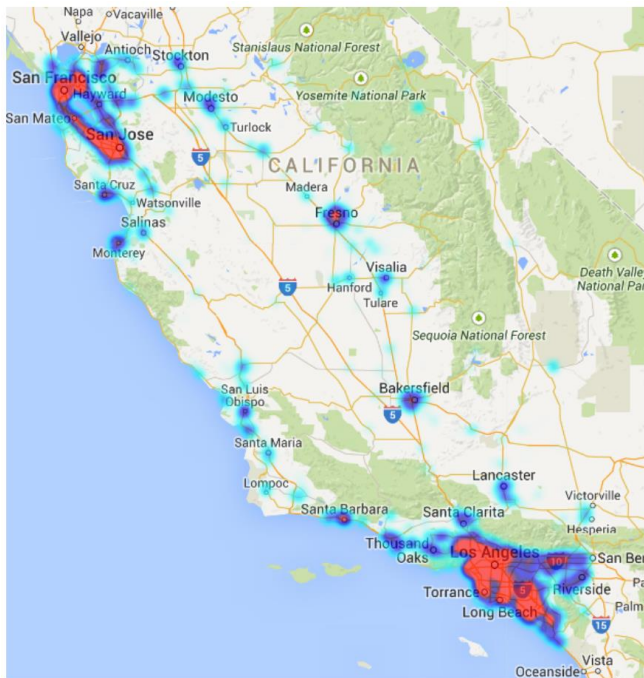


$$\widehat{r}_{ui} = \mu + b_u + b_i + (q_i + \frac{1}{|M(i)|} \sum_{n \in M(i)} |s_n|)^T p_u$$

set of geographic neighbours

impact of neighbours

Assignment 2



<i>"Fitness"</i>	<i>"Italian Restaurants"</i>	<i>"Airport & Rentals"</i>	<i>"Computer Repairs"</i>	<i>"Mexican"</i>
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

Assignment 2

Wikispeedia
navigation
traces:

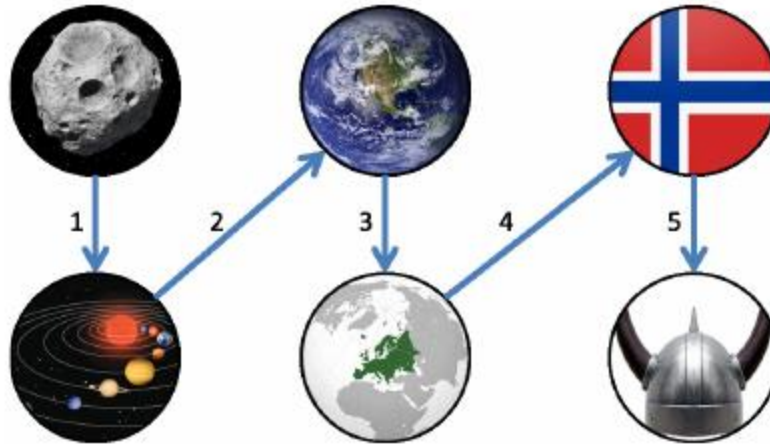


Figure 5: Graph of a complete path

	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

Assignment 2

Images from Chictopia



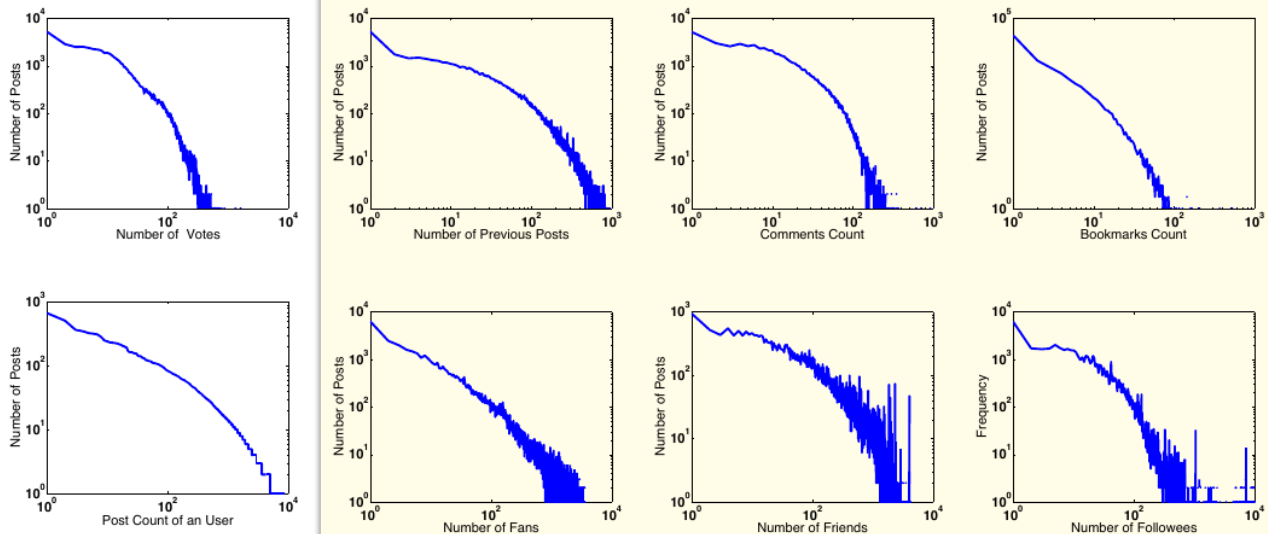
Tags electric, every day, summer, cute, T-shirt, chic

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!



Questions?