TweetStrap

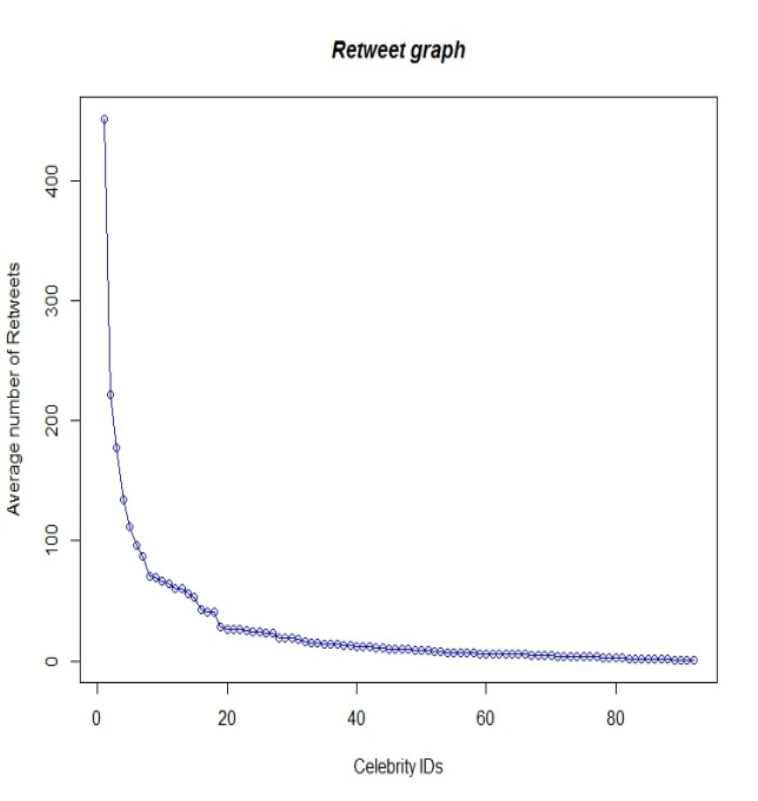
Natarajan Chakrapani
Kuldeep Kapade
Karthik Reddy Vadde

Twitter Mentor : Stanislav Nikolov

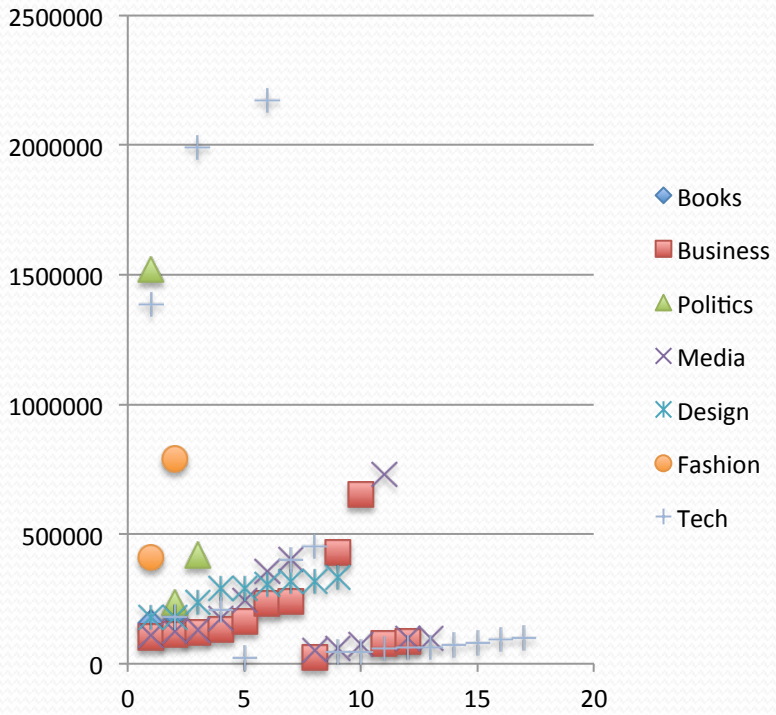
Goals

- "Tweetstrap" a marketer to identify mini-celebrities who can provide max retweets of their test tweets.
- Value : See beyond the usual suspects.
- Predict apriori retweet counts , if a test marketing tweet is tweeted by mini-celebrities from various domains
- Rank the celebrities by predicted retweet counts, based on their recent tweeting behavior

Avg Retweet count by celebrity



Followers by celebrity domains



Assumptions

- Tweets span topics
- Celebrities talk about different topics. Leverage that to identify tweets that might be easy to make them tweet about for max. reach
- Retweet count = Network reach = influence
- Retweets from immediate followers are significant in influence on global retweet counts compared to retweets from non-followers (via twitter search)

Data

- Extracted tweets and retweets of recommended tweeters for different categories.
  - (84 users across 14 categories)

- Verified accounts. Why?
  - 61 % of the tweets from these tweets get retweets.

Features Used : v1

- Tweet Features
  - Topic distribution of tweet (numeric vector of topic scores)
  - Topic Distribution of tweets aggregated by Author.

- Social Features
  - Local Retweet count from first level followers
  - Friend/Follower
  - Listed_count/Follower

Features Used : v2

- Tweet Features
  - Topic distribution of tweet (numeric vector of topic scores)
  - ~~Topic Distribution of tweets aggregated by Author.~~

- Social Features
  - Local Retweet count from first level followers
  - ~~Friend/Follower~~
  - ~~Listed_count/Follower~~

Topic Modeling

- Approach 1 : Latent Dirichlet Allocation (LDA) for topic modeling
- Approach 2 : use uclassify API for topic modeling.

- Use these topics as features for co-relating to retweet counts along with other user attributes.

Problems with LDA

- Tweet content too limited to suggest strong topics
- Semantically unrelated terms seen in topics
  - `2012-11-25 11:32:24,113 : INFO : topic #2: 0.016*out + 0.012*check + 0.008*help + 0.006*@ + 0.005*some + 0.005*celebrity + 0.005*one + 0.004*need + 0.004*me + 0.004*can + 0.004*love + 0.004*style + 0.004*$ + 0.004*news + 0.004*womens + 0.003*years + 0.003*th + 0.003*these + 0.003*do + 0.003*our`
- Hashtag terms overlap across topics
- Twitter spam hard to control

Uclassify

- A web service providing a public topic model classifier
- Built on top of the Open Directory project
- Multi level SVM Classification done over millions of documents
- Ten high level categories identified : (Arts, Business, Computers, Games, Health, Home, Recreation, Science, Society, Sports)

Approach 1

- Predict the retweet count given a text or topic with user as the centre using Multi-linear regression.

$$E(Y \mid X) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$$

Coefficients

Feature like # list count

- Goal: For a given tweet , Rank users based on predicted retweet counts.

Model 1:Multi Linear regression

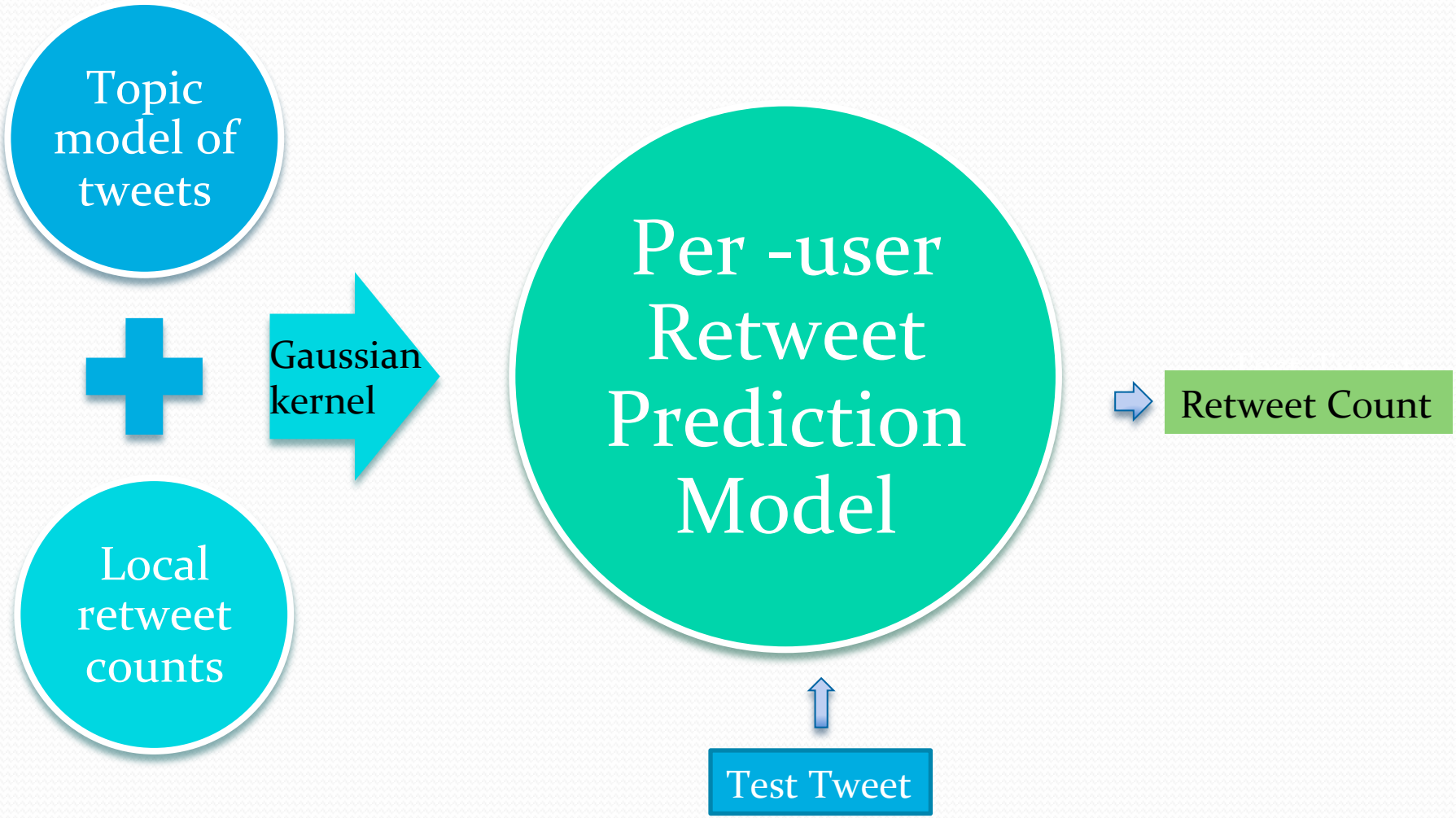
variable	coefficient	std. Error	t-statistic	prob.
const	-365.478526	1016.165069	-0.359665	0.719138
x1	452.602204	1024.552589	0.441756	0.658716
x2	445.124916	1026.104080	0.433801	0.664482
x3	379.665338	1024.600601	0.370550	0.711014
x4	427.162107	1023.273852	0.417447	0.676399
x5	438.522748	1024.964938	0.427842	0.668814
x6	437.297918	1024.153247	0.426985	0.669438
x7	406.350527	1024.402410	0.396671	0.691654
x8	468.469618	1026.450134	0.456398	0.648156
x9	420.150035	1025.998174	0.409504	0.682216
x10	480.435207	1027.406405	0.467619	0.640110
x11	-591.189392	1076.731942	-0.549059	0.583029
x12	-2279.465626	335.733960	-6.789500	0.000000
x13	947855.916033	26979.168879	35.132880	0.000000

Models stats		Residual stats	
R-squared	0.483405	Durbin-Watson stat	1.514446
Adjusted R-squared	0.479900	Omnibus stat	2662.954024
F-statistic	137.915497	Prob(Omnibus stat)	0.000000
Prob (F-statistic)	0.000000	JB stat	828587.886089
Log likelihood	-13696.033782	Prob(JB)	0.000000
AIC criterion	14.207289	Skew	7.695159
BIC criterion	14.247659	Kurtosis	103.333552

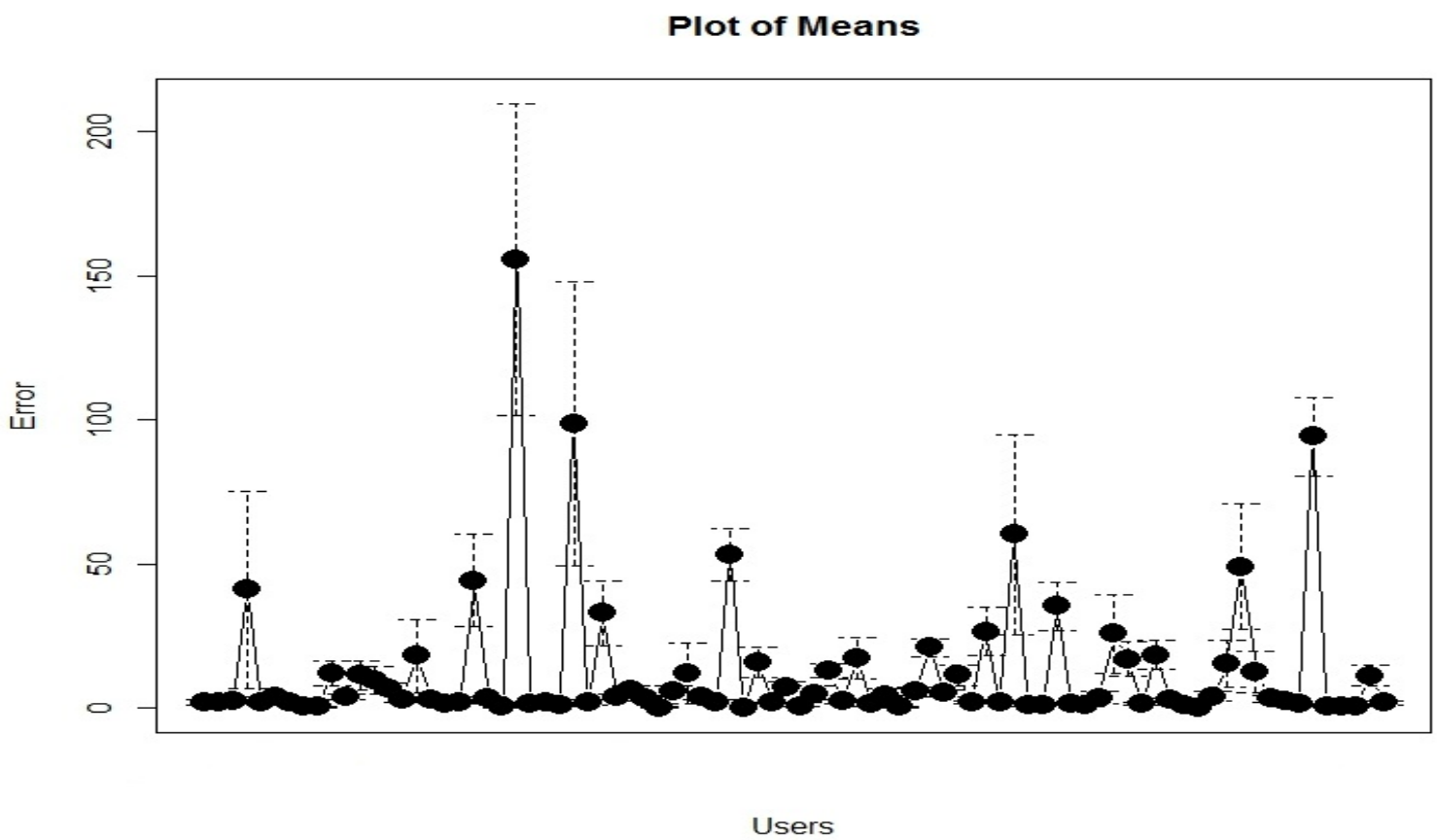
Approach 2 : Support Vector Regression

- Develop models per-user.
- Non linear relationship among the variables.
- For test tweet, estimate local retweet count based on retweet counts seen from "*similar*" tweets from user model.
- Use a Gaussian kernel (rbf) to estimate a best fit curve.
- Retweet_count = F(tweet topic scores , local retweet count)
- Compare the relative retweet counts across users

Prediction process



Prediction error margins



Demo

Questions??