

Influential People By Topic On Twitter

Girish Duvuru Raghavendra Madakkagari

Computer and Information Science and Engineering, University of Florida



INTRODUCTION

Word-of-mouth diffusion has long been regarded as an important mechanism by which information can reach large populations, possibly influencing public opinion, adoption of innovations, new product market share, or brand awareness. In recent years, interest among researchers and marketers alike has increasingly focused on whether or not diffusion can be maximized by seeding a piece of information or a new product with certain special individuals, often called “influentials” or simply “influencers,” who exhibit some combination of desirable attributes—whether personal attributes like credibility, expertise, or enthusiasm, or network attributes such as connectivity or centrality—that allows them to influence a disproportionately large number of others, possibly indirectly via a cascade of influence.

OBJECTIVE

Our objective is for a given topic X, find people who influence others most on twitter. A minority of members in a society possess qualities that make them exceptionally persuasive in spreading ideas to others. This concept is crucial in in sociology and viral marketing.

MATERIALS AND METHODS

Our Approach:

- Every user is an influencer and an influencee (i.e. can be influenced by somebody and can get influenced by somebody).
- Influence power depends not just on user influencing but depends on influencee’s interest on a topic.
- Influencer power is measured by user engagement of his interested influencees(followers) on a topic.

Effective followers count for a user x on topic t:

$I(x, t)$ = Interest of user x on topic t.

$0 \leq I(x, t) \leq 1$ and

$$\sum_{t \in Topics} I(x, t) = 1$$

$$EF(x, t) = \sum_{y \in followers(x)} I(y, t)$$

More EF implies more influential.

User Engagement:

$rc(x)$ = total retweets count of an user x

$mc(x)$ = total mentions count of an user x

$rc(x, t)$ = total retweet count of user x for topic t

$mc(x, t)$ = total mention count of user x for a topic t

$userEngagement(x) = (rc(x) + mc(x)) / \text{total tweets}$

$userEngagement(x, t) = (rc(x,t) + mc(x,t)) / \text{total tweets in topic t}$

Algorithm:

- Find users with more followers than friends. These are potential influential users.
- Topic model their tweets.
- Find most frequent topics.
- Model each of followers tweets and find the similarity measure w.r.t topic model (K-L divergence for distribution similarities).
- Gives us % of interest each user has in each of the topic
- For an Influential user find the sum of interest scores of each of followers to find Effective followers in each topic.

Data Used:

- Influential users: 780 users scraped from wefollow.com
- Bottle neck rate limit on api calls by twitter
- For each of 780 users 1000 tweets each (200 per api call)
- topic modeling on above using gensim
- For random 30 users among 780 we sampled 5000 followers for each of them
- For all 30*5000 users (api calls) we downloaded 200 tweets.

Implementation:

- Data cleaning.
- Remove: stop words, urls, mentions.
- Topic model for 780 follower tweets (corpus, num_topics).
- Trail and Error for No of topics chosen based on result (if we see duplicate topics reduce and merged topics increase).
- For 30 user’s followers find the similarity score w.r.t each topic and EF count for each topic.
- For each follower similarity score will return a vector $[t1, t2, t3, \dots, tn]$ ie score for each topic ($t1 + t2 + t3 + \dots + tn = 1$).
- For user find sum of followers topic vector to get EF.

Systems/Technologies used:

- Ruby scraping.
- D3.js for visualization
- Python topic model, dumbo.
- AWS computation.
- Tried EMR (with dumbo) to test distributed prototype for EF count.

Next Steps:

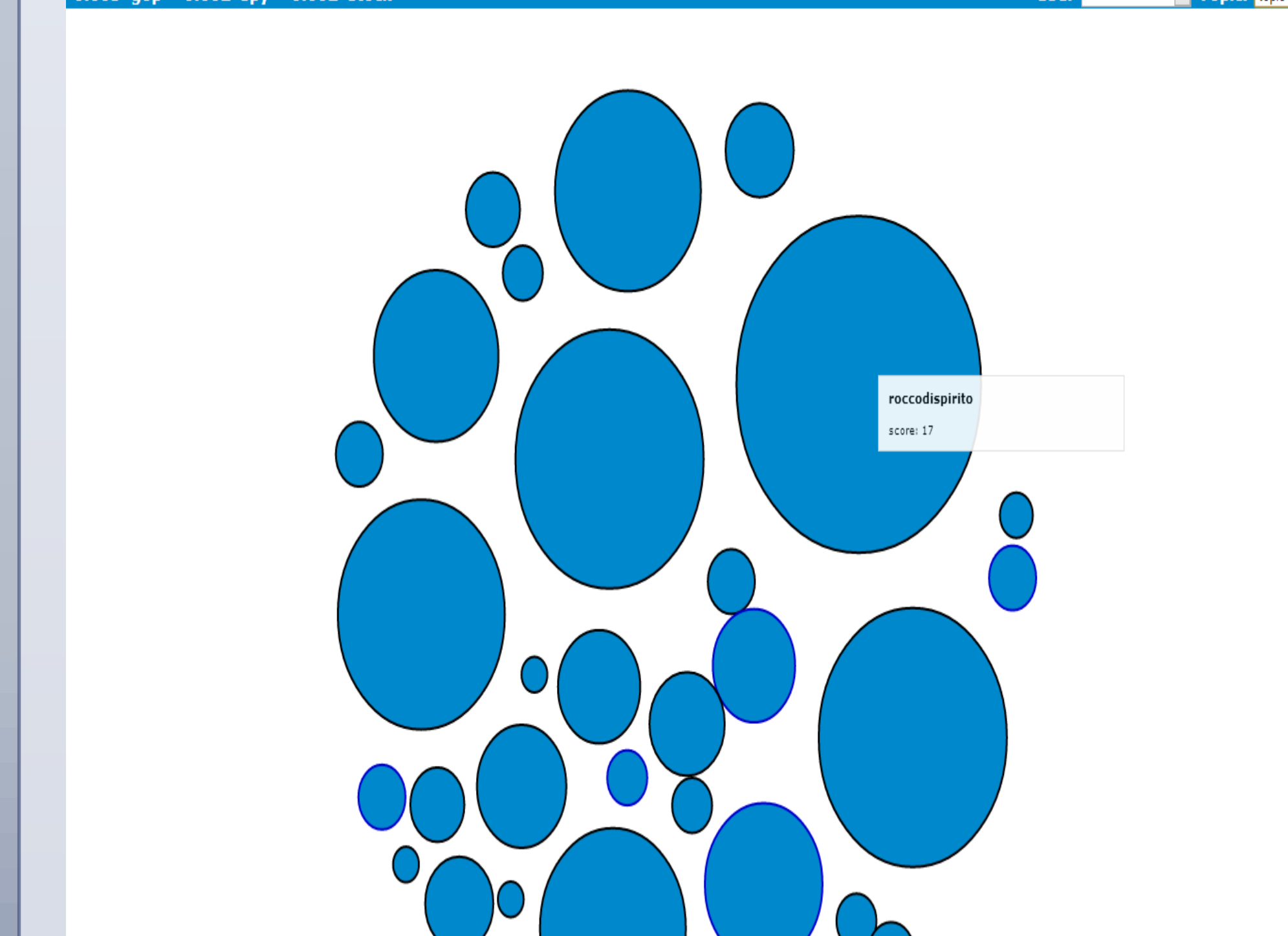
- Remove adjectives (great, awesome) topic mostly depend on nouns/verbs in our case (POS tagging).
- Results with more data.
- Page rank with influence score as initial values.

RESULTS

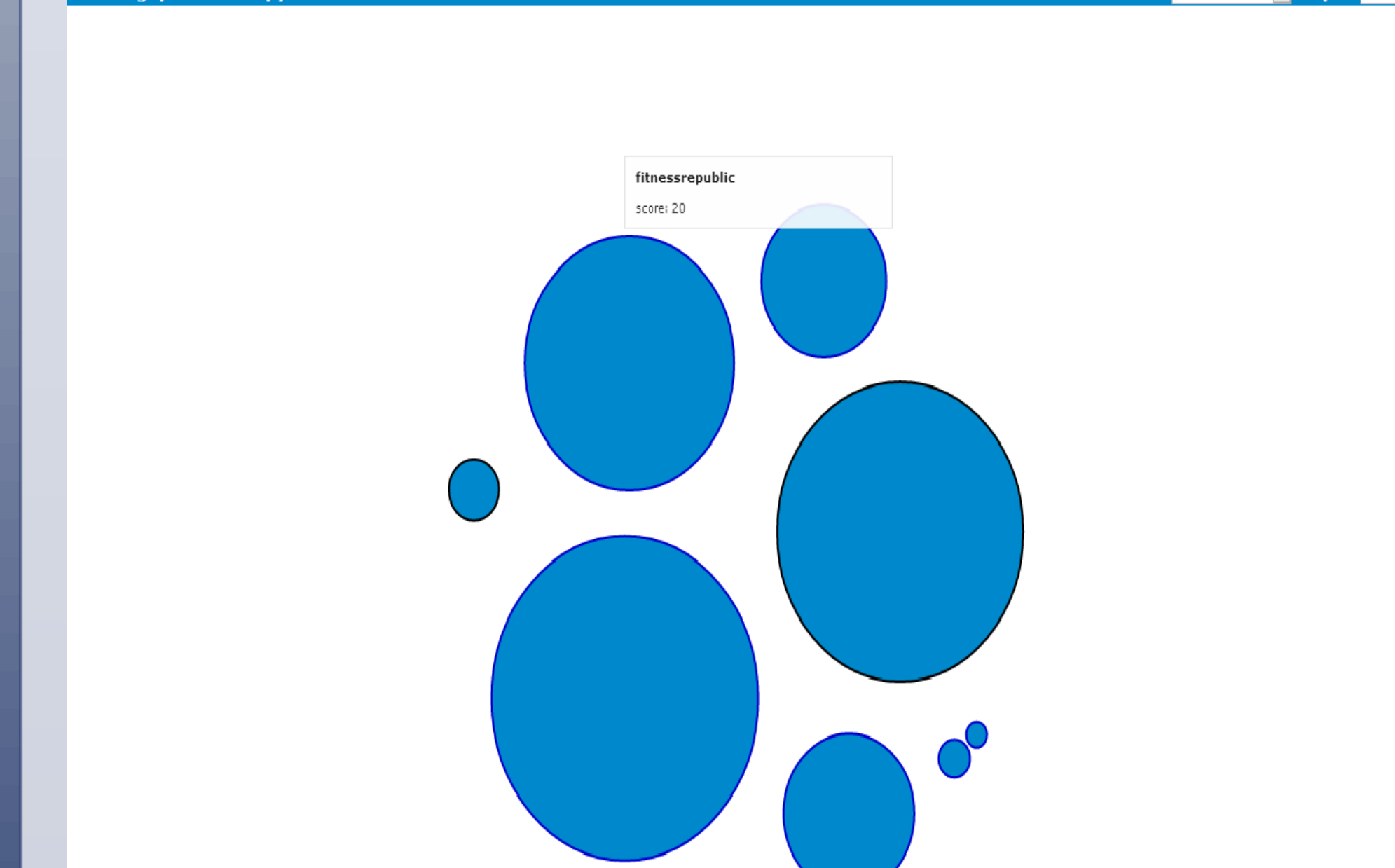
The results can be found at: <http://damp-sands-5714.herokuapp.com/> LDA for the topics:

topic 1 :0.007*obama + 0.005*stocks + 0.004*market + 0.003*aapl + 0.003*president + 0.003*watch + 0.003*trading + 0.003*gop + 0.002*spy + 0.002*stock
topic 2 :0.006*game + 0.003*team + 0.003*win + 0.003*show + 0.003*night + 0.003*people + 0.003*man + 0.003*nba + 0.002*play + 0.002*twitter
topic 3 :0.006*google + 0.004*apple + 0.004*money + 0.004*facebook + 0.003*people + 0.003*video + 0.003*app + 0.003*twitter + 0.003*social + 0.002*mobile
topic 4 :0.025*vegan + 0.006*food + 0.006*beauty + 0.004*recipe + 0.004*free + 0.003*blog + 0.003*eat + 0.002*recipes + 0.002*healthy + 0.002*animal
topic 5 :0.011*fashion + 0.007*nyfw + 0.006*show + 0.006*spring + 0.004*collection + 0.004*style + 0.004*beauty + 0.004*fall + 0.003*dress + 0.003*hair
topic 6 :0.013*food + 0.005*chef + 0.005*recipe + 0.004*wine + 0.004*restaurant + 0.003*chicken + 0.003*gold + 0.003*dinner + 0.003*eat + 0.002*recipes
topic 7 :0.022*travel + 0.005*world + 0.004*hotel + 0.003*trip + 0.003*tips + 0.003*food + 0.003*hear + 0.002*book + 0.002*ttot + 0.002*fun
topic 8 :0.018*health + 0.005*study + 0.004*risk + 0.004*cancer + 0.004*care + 0.003*news + 0.003*research + 0.003*heart + 0.003*embryo + 0.003*dr

Users visualization for topic 1 :0.007*obama + 0.005*stocks + 0.004*market + 0.003*aapl + 0.003*president + 0.003*watch + 0.003*trading + 0.003*gop + 0.002*spy + 0.002*stock



Users visualization for topic 1 :0.007*obama + 0.005*stocks + 0.004*market + 0.003*aapl + 0.003*president + 0.003*watch + 0.003*trading + 0.003*gop + 0.002*spy + 0.002*stock



In figure1: The circle with largest diameter shows the most influential user for that topic. In figure2: The circle with largest diameter shows the topic which is widely discussed by the followers of that user.

CONCLUSIONS AND LESSONS LEARNT

From the analysis we did on the twitter data we were able to conclude that users with more followers are generally very focused. They generally talk about 1.7 topics on an average. General users usually focus on 5 topics on an average.

For ranking users in a specific topic, modeling by user interests and the method of effective followers could be a great start for page rank algorithms.

Also we were able to learn some lessons from this project.

- Cleaning is most important part in data analysis. Without removing **twitter specific stop words** (mentions, lols, thnx) topics are meaningless.
- Compression saves lot of space and IO time (our case 1.1 G to 52 M).
- Also save data in one large file (document per line in our case) instead of many small files. Don’t need to open many files and copies would be faster (We downloaded in 6 micro instances and moved to main computation machine).

REFERENCES

- Bakshy E, Hofman M J, Wason A. M, Watts J. D. 2011. Everyone’s an Influencer: Quantifying Influence on Twitter. WSDM
- Sasa Petrovic , Miles Osborne , Victor Lavrenko. 2011. RT to Win! Predicting Message Propagation in Twitter. International Conference on Weblogs and Social Media (ICWSM).
- Roja Bandari, Sitaram Asury and Bernardo Hubermanz. The Pulse of News in Social Media: Forecasting Popularity.

ACKNOWLEDGEMENTS

Dr. Daisy Zhe Wang
Computer and Information Science and Engineering,
University of Florida

CONTACTS

Girish Duvuru
University of Florida
girish.iiith@gmail.com

Raghavendra Madakkagari
University of Florida
raghu.snist@gmail.com