CS 5604 Informational Storage & Retrieval Spring 2015
Social Networks & Importance

04/30/2015

Bharadwaj Bulusu - bbsb08@vt.edu
Vanessa Cedeno- vcedeno@vt.edu
Islam Harb - iharb@vt.edu
Yilong Jin – jin28@vt.edu
Sai Ravi Kiran Mallampati - sairavi5@vt.edu

Social Network and Importance

- **Our Role (SN Team)**
  - Pick up the social non-content based features .
  - Calculate the Importance Values for Tweets and Web pages based on social network features and connectivity.

Non-Content Features

- Content-based features are out of scope.
- Features are considered with
  - Connectivity (Popularity)
  - Trustworthiness
- This metric “Trustworthiness and popularity” will be mainly reflected in the “User Importance Value (UIV)”.

- **Tweet Specific Features**
  - Retweet count: The number of times the tweet has been retweeted.
  - Favorites count: The number of times the tweet has been favored.
- **Account authority (User) features**
  - Followers count: The number of followers a user has.
  - List Count: The number of lists a user belongs to.

UIV: User Importance Value

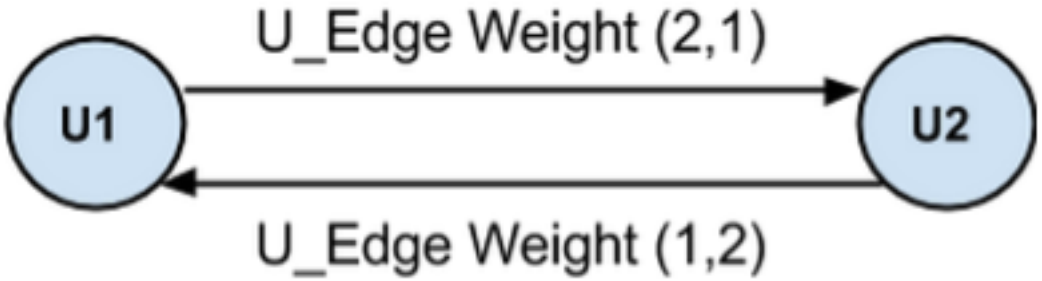


Figure 1: User nodes

U_Edge Weight(1,2) = (#Mentions of U1 made by U2) / #Mentions[U2] .... Equation (1)

U_Edge Weight(2,1) = (#Mentions of U2 made by U1) / #Mentions[U1] .... Equation (2)

Where:

#Mentions[U1] : Total number of tweets posted by U1 that mention other users.

#Mentions[U2] : Total number of tweets posted by U2 that mention other users.

$$UIV = \begin{cases} \Sigma\,(U\_Edge\ Weight) \ / \text{ number of inlink edges} \\ 0 \quad \text{, if there are no inlink edges} \end{cases}$$

Tweet Importance Value

$$TIV = \sum_{i=1}^{5} W(i) * A(i) \,/\, Number\ of\ attributes$$

Weights are arbitrary.

-For Simplicity, chosen to be uniform.

- E.g. W1 = W2 = W3 = W4 = W5 = 0.2.

The Five Attributes

-A1: Favorite Count

(# Fav(i) - Fav(min)) / (Fav(max) - Fav(min))

-A2: Retweet Count

(# RT(i) - RT(min)) / (RT(max) - RT(min))

-A3: List Count

(# List(i) - List(min)) / (List(max) - List(min))

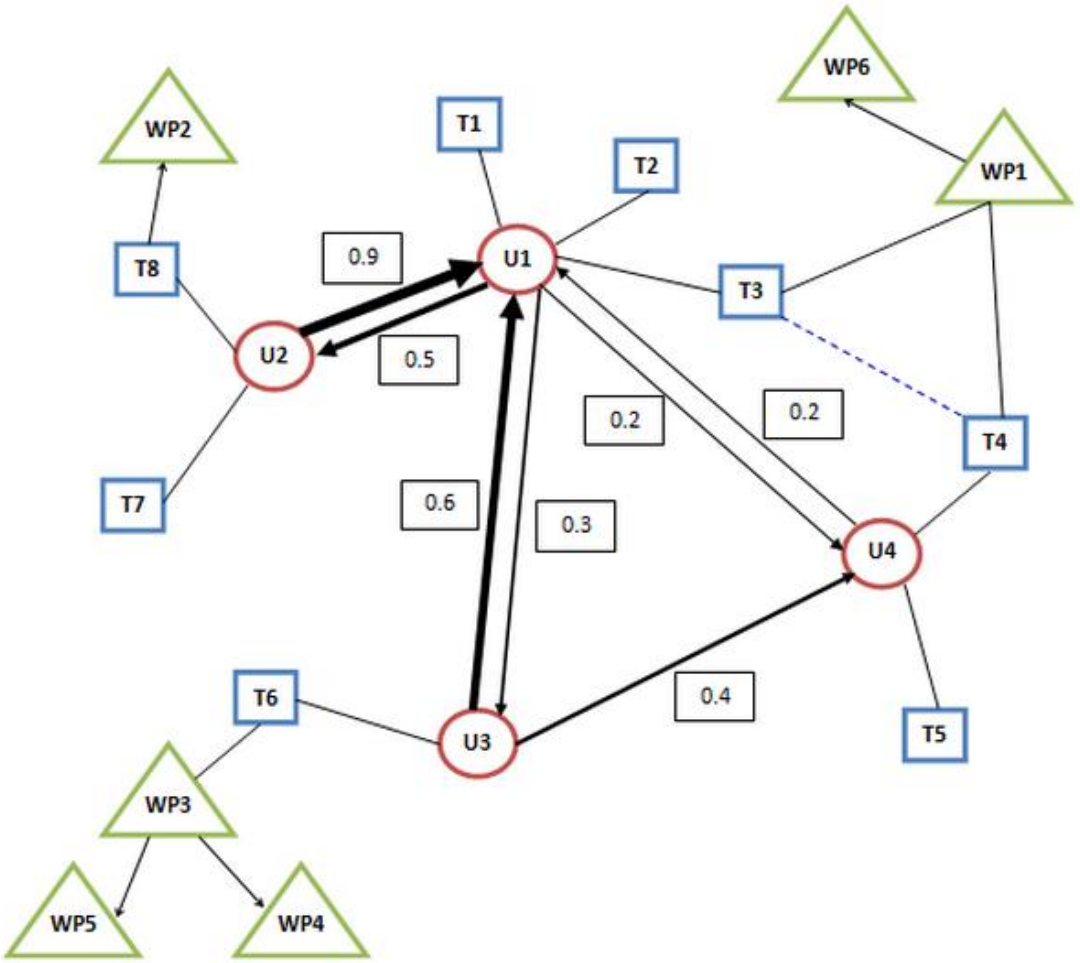
-A4: Number of Followers

(# Followers(i) - Followers(min)) / (Followers(max) - Followers(min))

- A5: User Importance Value (UIV)

The graph

- **Users**:
  - Edge indicates a mention occurred.
- **Tweets**:
- Dotted Edges between tweets indicates a retweet.
- Solid Edges between Tweet and a user indicates the tweet's ownership.
- **Web pages**:
  - Edges indicates a link/connectivity.

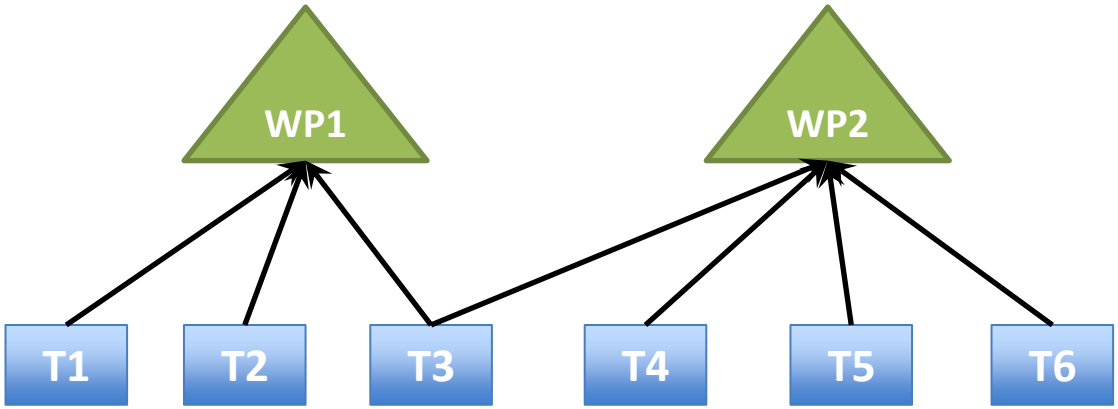


Webpage Importance Value

$$WPIV_j = \frac{\sum TIV(i)}{t} * \frac{t - Minimum\ number\ of\ edges\ to\ a\ webpage + 1}{Maximum\ number\ of\ edges\ to\ a\ webpage - Minimum\ number\ of\ edges\ to\ a\ webpage + 1}$$

where,

- "t" is the number of tweets that points to WPIV(j).
- Example:
  - For WP1, the "t" is equal to 3 (comes from T1->T3).
  - For WP2, the "t" is equal to 4 (comes from T3->T4).



Approach (Data Structure)

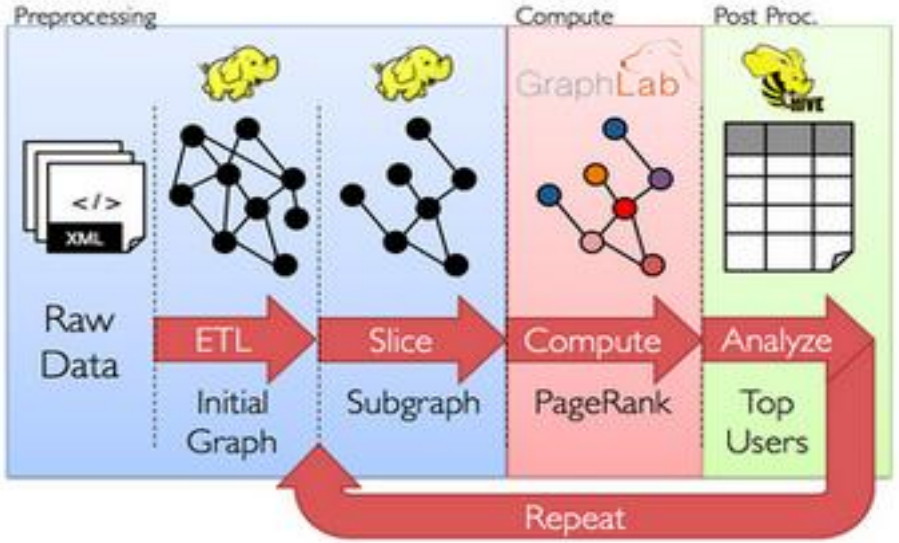
- JSON File
  - Avro Tweets : Comply with other teams schema
  - Avro User: Only for SN team, basically
    - Number of Followers
    - List Count.

Approach (Sequential vs. Parallel)

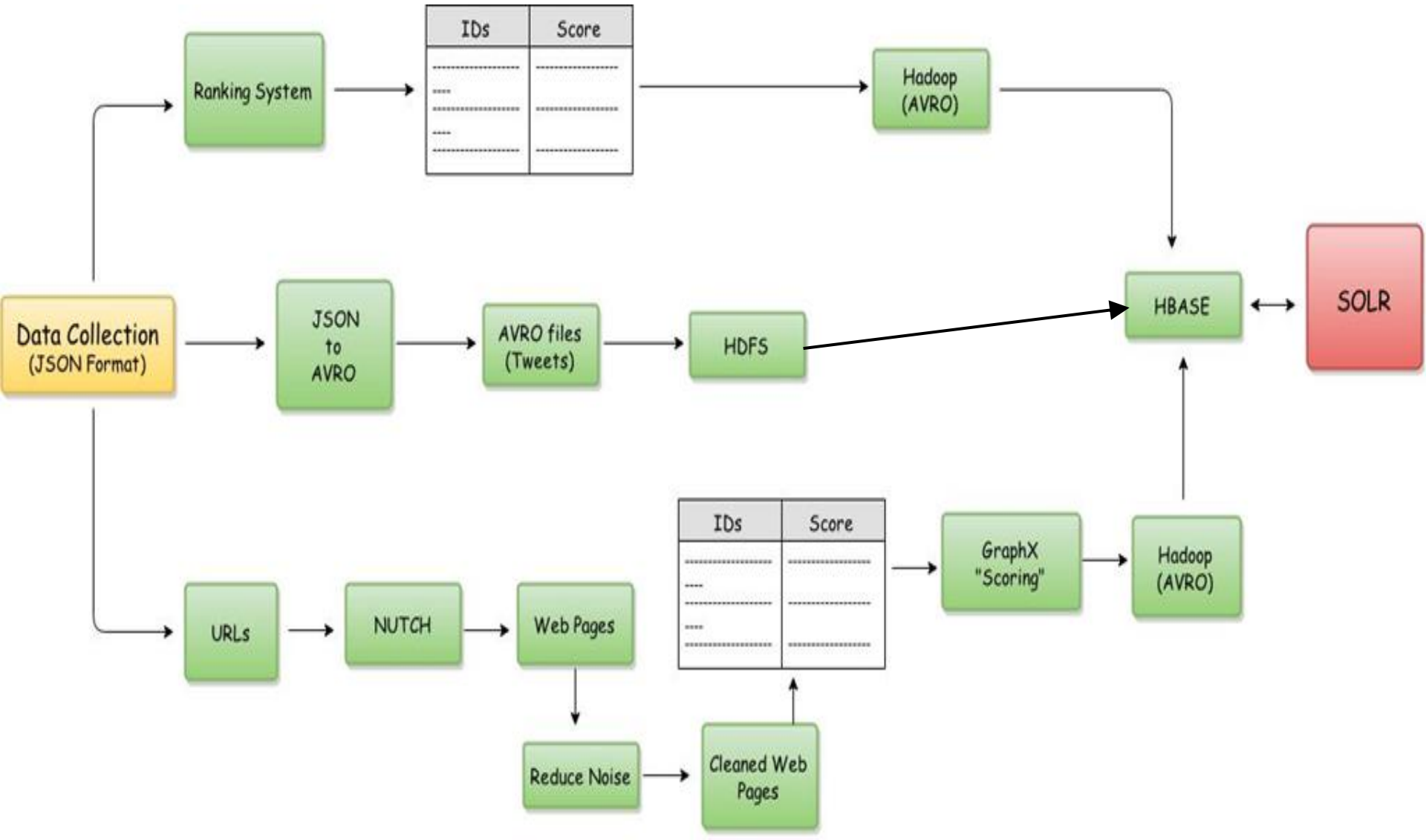
- Parallelizing the “Parser” in Python
  - Using Standard Python Multi-processing library.
  - Local machine (Jin’s) – 8 cores
  - ~5x improvement/speedup

GraphX

- Introduces the Resilient Distributed Property Graph: a directed multigraph with properties attached to each vertex and edge.
- RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators.
- Includes a growing collection of graph algorithms and builders to simplify graph analytics tasks



Project Model



Statistics

- Very small data collection
  - 2553 tweets
  - num of user: 915
  - Reading tweets time: 0.612 second(s)
  - Reading User time: 0.066 second(s)
  - Updating Edges: 0.012 second(s)
  - Calculating UIV: 0.023 second(s) // this contains users that are not in the dataset
  - Calculating tweet score: 0.048 second(s)

Statistics

- Small data collection
  - 428574 tweets
  - 265410 users
  - Reading tweets time: 110.551 second(s)
  - Reading User time: 22.568 second(s)
  - Updating Edges: 3.821 second(s)
  - Calculating UIV: 0.951 second(s) // this contains users that are not in the dataset
  - Calculating tweet score: 6.383 second(s)

References

1. Events Archive. Project History Overview. Retrieved 16:30, February 11, 2015, from http://www.eventsarchive.org/?q=node/70

2. Tianyi Wang et al., 2011. "Understanding Graph Sampling Algorithms for Social Network Analysis", 31st International Conference on Distributed Computing Systems Workshops (ICDCSW), Pages 123-128, Minneapolis, MN, USA.

3. How we analyzed Twitter social media networks with NodeXL, http://www.pewinternet.org/files/2014/02/How-we-analyzed-Twitter-social-media-networks.pdf

4. Seth A. Myers et al., 2014. Information Network or Social Network?:The Structure of the Twitter Follow Graph, Proceedings of the companion publication of the 23rd international conference on World Wide Web companion, Pages 493-498, ISBN: 978-1-4503-2745-9, Geneva, Switzerland.

5. Srijith Ravikumar, Raju Balakrishnan, and Subbarao Kambhampati. 2012. Ranking tweets considering trust and relevance. In Proceedings of the Ninth International Workshop on Information Integration on the Web (IIWeb '12). ACM, New York, NY, USA, pages 4. http://doi.acm.org/10.1145/2331801.2331805

6. Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. 2010. An empirical study on learning to rank of tweets. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 295-303. http://dl.acm.org/citation.cfm?id=1873815

7. Serge Abiteboul, Milhai Preda, Gregory Cobena. 2003. Adaptive On-line Page Importance Computation, Proceedings of the companion publication of the 12th international conference on World Wide Web companion, Pages 280-290, Budapest, Hungary. http://dl.acm.org.ezproxy.lib.vt.edu/citation.cfm?id=775152

8. Patrick Dlogan, python-json 3.4, http://sourceforge.net/projects/json-py/. Accessed 2-10-2015.

9. John Hunter, matplotlib, http://matplotlib.org. Accessed 2-12-2015.

10. NetworkX, https://networkx.github.io. Accessed 2-12-2015.

11. Graphviz - Graph Visualization Software, retrieved from http://www.graphviz.org. Accessed 2-13-2015.

12. Solr in 5 minutes, retrieved from http://www.solrtutorial.com/solr-in-5-minutes.html. Accessed 1-26-2015.

13. Big Data and Hadoop. April 21, 2014. Installing Hadoop on Mac OS X Mountain Lion 10.8.5 bit. Retrieved 11:00, February 20, 2015 from http://glebche.appspot.com/static/hadoop-ecosystem/hadoop-hive-tutorial.html#hadoop-installation-mac-osx

14. Spark. Quick Start. Retrieved 15:00, February 20, 2015 from https://spark.apache.org/docs/latest/quick-start.html

References

15. Ampcamp. GraphX. Retrieved 10:00, February 23, 2015 from http://ampcamp.berkeley.edu/big-data-mini-course/graph-analytics-with-graphx.html#constructing-an-end-to-end-graph-analytics-pipeline-on-real-data

16. Apache Giraph. Quickstart. Retrieved 15:00, February 23, 2015 from http://giraph.apache.org/quick_start.html

17. Hadoop application architectures. Chapter 4. Graph processing on Hadoop. Retrieved 17:00, February 23, 2015 from https://www.safaribooksonline.com/library/view/hadoop-application-architectures/9781491910313/ch04.html

18. Apache Giraph. Introduction. Retrieved 18:00, February 23, 2015 from http://giraph.apache.org/intro.html

19. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The million follower fallacy. In Proceedings of the Association for the Advancement of Artificial Intelligence. USA.

20. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media? Pages 591-600. In Proceedings of the 19th international conference on World Wide Web.

21. Eytan Bakshy, Winter A. Mason, Jake M. Hofman and Duncan J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. 2011. In Proceedings of the ACM WSDM 2011. Pages 65-74. Hong Kong.

22. The Engineering Behind Twitter's New Search Experience. By Twitter Search. Retrieved 10:00, March 10, 2015 from https://blog.twitter.com/2011/engineering-behind-twitter%E2%80%99s-new-search-experience

23. Hang Li. 2011. A Short Introduction to Learning to Rank. IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10

24. Page Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University.

25. Solrpy project Retrieved March 22, 2015 from https://code.google.com/p/solrpy/

26. Nutch Tutorial. Introduction. Retrieved 20:00, March 27, 2015 from https://wiki.apache.org/nutch/NutchTutorial

27. Nutch Tutorial. Retrieved 14:00, March 23, 2015 from https://scholar.vt.edu/access/content/group/5508d3d6-c97d-437f-a09d-2cfd43828a9d/Tutorials/Nutch%20Tutorial.pdf

28. Reading and Writing Avro Files From the Command Line, Retrieved March 28, 2015 from http://www.michael-noll.com/blog/2013/03/17/reading-and-writing-avro-files-from-the-command-line/


ngiyabonga
danke
謝謝
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
mauruuru
mochchakkeram
dziękuje
go raibh maith agat
obrigado
sukriya
kop khun krap
grazie
arigato
dakujem
merci
감사합니다

Appendix A

top 3 tweets

Israel will be allowed to "act defensively" on tunnels during cease-fire, U.S. says.

http://t.co/4sxohD3iXV

cnnbrk

score 0.240

UIV: None

# of followers: 17842891

list count: 155404

mentioned by users in the collection 0 times

retweet count: 0

fav_count: 0

Appendix A

Ebola vaccine human tests could begin as early as September, National Institutes of Health

says. http://t.co/dYwYaqnsf7

cnnbrk

score 0.240

UIV: None

# of followers: 17842891

list count: 155404

mentioned by users in the collection 0 times

retweet count: 0

fav_count: 0

Appendix A

RT @JustCallMe_Mese: Retweeting this literally takes 1.8 seconds http://t.co/7MV02iwKVj

miketanz11

score 0.200

UIV: 0

# of followers: 221

list count: 0

mentioned by users in the collection 0 times

retweet count: 55897

fav_count: 0
