Analyzing Influence Metrics in Twitter

NCSR Demokritos

Excercise for the Reveal Project

Marinos Galiatsatos
February 24, 2016

Contents

Introduction

In this test we create four ranking lists from a graph that represents tweets . More specifically, we create a degree, betweenness, pagerank and followers ranking lists and apply to them the Kendall's Tau correlation. Also, we try to find similarities among the tweet's content of the top users of each ranking list.

Problem

Suppose you have a stream of data which is produced by a set of users and you want to find the most influential users in this set. If you represent the data as a graph where the nodes are the users and the edges are the interactions between the users then your problem can be translated in finding the most central nodes in the graph.

Implementation

For the implementation we used Python, with the NetworkX package for the creation of the Graph, and sklearn package for the normalization of the rank values. Other usefull packages that were used are scipy, pandas and json.

First, we load the json file. For the Graph Nodes we collect all the users with their details and the hashtags that they may have used in the tweet. For the Graph Edges we combine the users with the mention users for a tweet. If a user A has mentioned B and C, then we create 2 edges (A-B) and (A-C). Also, because some users don't use hashtags we store the tweet text, in order to use it later for better results in content similarities.

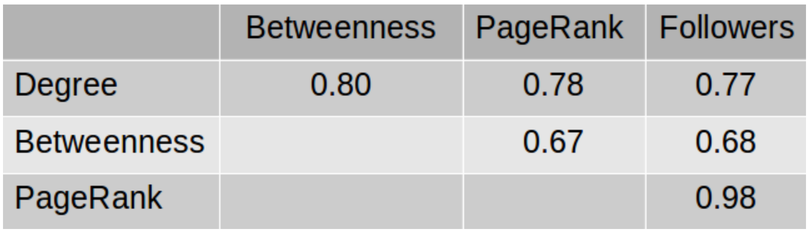
Second, after the creation of the Graph, we apply 3 centrality measures on it, degree, betweenness and pagerank, and we create 3 ranks. We also create a rank of the users with respect to the number of followers that they have.

Finally, we created 4 more lists with the top 5 users from every ranking list. Using the similar function that python provides, we compare the tweets among the users from each ranking list. Similar function returns a value between 0 and 1 and finds the relation between 2 strings. For example, similar(“Apple”, “Appel”)=0.8 but similar (“Apple”, “Mango”)=0.0.

Results

Kendall's Tau correlation

We now compare all the rankings with the Kendall's Tau correlation metric. This metric compares two rankings and returns a value between -1 and 1. Values close to 1 indicate that the rankings have strong aggrement and values close to -1 strong disaggrement. The table below shows the comparisons between all 4 rankings:



	Betweenness	PageRank	Followers
Degree	0.80	0.78	0.77
Betweenness		0.67	0.68
PageRank			0.98

As we can see from the table, the strongest correlation is between PageRank ranking and Followers ranking. A node (user) has high pagerank if he/she has many incoming links. Also, if a node (user) has many followers, then in a graph this is represented by incoming links. So we can say that if a user has many followers has also a high pagerank.

Another observation that we could make is that betweenness ranking with pagerank and followers rankings does not seem to have strong correlation in comparison to the other pairs.

Finally, we could say that degree has strong correlation with all the other rankings, because degree measures all the incoming and outgoing links from a node. So pagerank and followers represent the incoming links and betweenness represents the sum of the incoming and outgoing links.

Content Similarity

We compared the top 5 ranks from every ranking list, based on the tweet content and the hashtags.

Degree ranks:

Users' id	keywords
1 : 73398952	RT, Free #Venezuela, #Ukraine is with you!, #euromaidan
2 : 65167510	RT, Free #Venezuela, #Ukraine is with you!, #euromaidan
3 : 512700138	RT, Fight for the right to be free!! Fight Fascism everywhere! Free Venezuela the Ukraine And Russia
4 : 87818409	RT, Geoengineering side effects could be potentially disastrous
5 : 428333	RT, United States expels Venezuelan diplomats in tit-for-tat

user	user	Content similarity
1	2	1.0
1	3	0.40145985401459855
1	4	0.2857142857142857
1	5	0.39111111111111113
2	3	0.40145985401459855
2	4	0.2857142857142857
2	5	0.39111111111111113
3	4	0.3018867924528302
3	5	0.42424242424242425

4	5	0.28703703703703703

Betweenness ranks:

Users' id	keywords
1 : 87818409	RT, Geoengineering side effects could be potentially disastrous, research shows
2 : 73398952	RT, Free #Venezuela, #Ukraine is with you!, #euromaidan
3 : 428333	RT, United States expels Venezuelan diplomats in tit-for-tat
4 : 1652541	#Lululemon, #Athleta, RT Bitcoin exchange Mt. Gox goes dark in blow to virtual currency
5 : 2467791	RT, You can be put to death for homosexual acts in these countries

user	user	Content similarity
1	2	0.2625482625482625
1	3	0.28703703703703703
1	4	0.24896265560165975
1	5	0.36363636363636365
2	3	0.39111111111111113
2	4	0.224
2	5	0.2824427480916031
3	4	0.37681159420289856
3	5	0.3470319634703196
4	5	0.319672131147541

PageRank ranks:

Users' Id	keywords
1 : 73398952	RT, Free #Venezuela, #Ukraine is with you!, #euromaidan
2 : 512700138	RT, Fight for the right to be free!! Fight Fascism everywhere! Free Venezuela the Ukraine And Russia
3 : 65167510	RT, Free #Venezuela ! #Ukraine is with you! #euromaidan
4 : 18948541	RT, https//...
5 : 428333	RT, United States expels Venezuelan diplomats in tit-for-tat

user	user	Content similarity
1	2	0.40145985401459855
1	3	1.0

1	4	0.2840909090909091
1	5	0.3911111111111113
2	3	0.4014598540145985
2	4	0.26373626373626374
2	5	0.42424242424242425
3	4	0.2840909090909091
3	5	0.3911111111111113
4	5	0.3308270676691729

Followers ranks:

user	user	Content similarity
1	2	0.14883720930232558
1	3	0.10294117647058823
1	4	0.2
1	5	0.200836820083682
2	3	0.14184397163120568
2	4	0.1837837837837838
2	5	0.16393442622950818
3	4	0.18867924528301888
3	5	0.19393939393939394
4	5	0.3827751196172249

Users' Id	keywords
1 : 759251	@CNN @BBCNews @RT_com @maddow @InfowarsFeed ??????? 7 times Sir. That's certain....Are you certain Sir?
2 : 15485441	#FallonTonight I'm in. Lake Michigan will have a New York fish this weekend. #SwimmyFallon
3 : 21425125	@iamthereallhud @Aloiye_ :) *AM
4 : 807095	RT, 3 NYT stories you should know about now, #NoMatterWhat
5 : 18220175	This is what happens when you hit the court with @AQUAhydrate!! Bring your A game!! ,#NoMatterWhat

As we can see from the above results the degree, betweenness and pagerank ranking lists have tweets that have related content. On the other hand, the followers ranking list shows that the tweets aren't related with each other except for 2.