Homework # 2
Information Retrieval CSCE 670
Spring 2013

This File contains 4 parts each describing the Parts of the homework respectively. There is also a description of Bonus part implementation and results at last.
For Zip folder contains: 6 items:

1.part1.py : TF-IDF
2.part2.py : PageRank
3.part3.py : TF-IDF and PageRank Combined
4.bonus.py : Contains bonus part implementation
5.tweet.py : its just a class file which is already included in each file, so you don't have to do anything with it, but should be present in the same folder.

Query I have choosed:

1. **marscuriosity landing live on times square**
2. **time for live hangout of msl landing**
3. **awesome view of its landing**

Part1: Tf-Idf Calculation

Run: python part1.py

Used tf-idf to build vector space model. Each tweet text can be considered as a document
Query:

********* Enter Query To Search ***********

marscuriosity landing live on times square

*************************** RESULT ****************************

[Rank 0] : Tweet: [@marscuriosity at times square!]
[Rank 1] : Tweet: [on my way to times square for the #curiosity landing.]
[Rank 2] : Tweet: [watching the #curiosity landing from times square. because i can.]
[Rank 3] : Tweet: [kinda want to go to times square to watch the live broadcast of the mars landing. but that would involve going to times square.]
[Rank 4] : Tweet: [anyone at the @marscuriosity viewing in times square?]
[Rank 5] : Tweet: [wishing @marscuriosity the best as we watch the #historic #mars landing from times square! (@ times square) [pic]: http://t.co/cmgmvtfa]

[Rank 6] : Tweet: [times square for #marscuriosity? yeah, i'm that geek.]
[Rank 7] : Tweet: [#planetfest just went live on video to times square #msl]
[Rank 8] : Tweet: [i wish i were at times square right now... #msl #curiosity]
[Rank 9] : Tweet: [. @nasa in times square! #msl http://t.co/1mvcg5sw]
[Rank 10] : Tweet: [live coverage of the #msl landing is running on the big screen in times square!]
[Rank 11] : Tweet: [heading off to times square to cover the #curiosity landing.]
[Rank 12] : Tweet: [does anyone have a pic from times square #msl?]
[Rank 13] : Tweet: [being at times square tonight to watch the @marscuriosity land would have been awesome!]
[Rank 14] : Tweet: [nothing but times square on tv for mars landing? how depressing]
[Rank 15] : Tweet: [nasatv is streaming live on the main screen in times square. #msl]
[Rank 16] : Tweet: [heading to times square to watch us try and land this thing on mars.]
[Rank 17] : Tweet: [waiting for curiosity @ times square http://t.co/1abgeplx]
[Rank 18] : Tweet: [nice. the mars curiosity landing is going to be on the main screen in times square]
[Rank 19] : Tweet: [@starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 20] : Tweet: [wish i was in nyc tonight so i could go to times square and watch the curiosity landing]
[Rank 21] : Tweet: [rt @morningedition: anyone at the @marscuriosity viewing in times square?]
[Rank 22] : Tweet: [just saw @lori_garver on the big screen in times square! #msl]
[Rank 23] : Tweet: [3 hours to go for curiosity. who's going to times square to watch it on the big screen?]
[Rank 24] : Tweet: [if you're in nyc you can watch the mars rover landing in times square on the big screen!]
[Rank 25] : Tweet: [heading down to times square voluntarily to watch the mars rover landing. #nasa]
[Rank 26] : Tweet: [mars landing @ times sq. get going rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/bwkpr5xs]
[Rank 27] : Tweet: [at times square. about fifty people here, at best guess. #marscuriosity]
[Rank 28] : Tweet: [rt @sarabec: . @nasa in times square! #msl http://t.co/1mvcg5sw]
[Rank 29] : Tweet: [rt @sarabec: . @nasa in times square! #msl http://t.co/1mvcg5sw]
[Rank 30] : Tweet: [rt @sarabec: . @nasa in times square! #msl http://t.co/1mvcg5sw]
[Rank 31] : Tweet: [rt @sarabec: . @nasa in times square! #msl http://t.co/1mvcg5sw]
[Rank 32] : Tweet: [fingers crossed! (@ nasa curiosity mars mission broadcast in times square) http://t.co/trl23a3f]
[Rank 33] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 34] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 35] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 36] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 37] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 38] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 39] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]

[Rank 40] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 41] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 42] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 43] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 44] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 45] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 46] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 47] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 48] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[Rank 49] : Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]

*********Enter Query to search********

time for live hangout of msl landing

********** RESULT **********

[Rank 0] : Tweet: [time for the curiosity landing! #nasa]
[Rank 1] : Tweet: [#msl time!]
[Rank 2] : Tweet: [nerd time: @nasa feed, @badastronomer g+ hangout, and the real-time animation for the @marscuriosity landing.]
[Rank 3] : Tweet: [time to watch the landing of curiosity!]
[Rank 4] : Tweet: [mars time!]
[Rank 5] : Tweet: [mars time]
[Rank 6] : Tweet: [time to watch the curiosity landing.]
[Rank 7] : Tweet: [mars rover time.]
[Rank 8] : Tweet: [looking for a place to hangout with other space tweeps tonight? join us tonight! look for the g+ hangout link a little later. #msl #edl]
[Rank 9] : Tweet: [looking for a place to hangout with other space tweeps tonight? join us tonight! look for the g+ hangout link a little later. #msl #edl]
[Rank 10] : Tweet: [time for bruno mars.]
[Rank 11] : Tweet: [@marscuriosity is there a link to the google hangout?]
[Rank 12] : Tweet: [rt @anthonyfitch: looking for a place to hangout with other space tweeps tonight? join us tonight! look for the g+ hangout link a little ...]
[Rank 13] : Tweet: [curiosity en marte: cobertura del aterrizaje desde google+ hangout // google+ hangout — curiosity landing coverage http://t.co/3b1jj6z4]

[Rank 14] : Tweet: [excited and watching the g+ hangout coverage for the curiosity landing! #spacenerd]
[Rank 15] : Tweet: [. @nasa_tv now on big screen, @cosmoquestx g+ hangout on laptop, holding for live hangout on pad. #msl #nosleeptilmars]
[Rank 16] : Tweet: [it's just about time for the mars landing]
[Rank 17] : Tweet: [watching the curiosity google+ hangout. yay, space!]
[Rank 18] : Tweet: [curiosity mars landing g+ hangout is live! http://t.co/roz8ep3a]
[Rank 19] : Tweet: [watching the mars curiosity hangout http://t.co/2tbdgxol #marscuriosity]
[Rank 20] : Tweet: [google+ hangout — curiosity landing coverage #live http://t.co/ihmn0exd]
[Rank 21] : Tweet: [go time #curiosity]
[Rank 22] : Tweet: [curiosity mars landing g+ hangout is live! http://t.co/roz8ep3a #marshangout]
[Rank 23] : Tweet: [google+ hangout covering the #msl curiosity landing is now live. http://t.co/7nioc65n]
[Rank 24] : Tweet: [yay! we got into the #hangout!! rt @universetoday "you can watch our live virtual landing party for #curiosity." #marshangout #msl]
[Rank 25] : Tweet: [watch the curiosity hangout here! http://t.co/rmviuazm]
[Rank 26] : Tweet: [watch the curiosity hangout here! http://t.co/2ovlhwim]
[Rank 27] : Tweet: [mars curiosity landing hangout happening now.. http://t.co/obpqd7ih]
[Rank 28] : Tweet: [we are now live!! google+ hangout covering the full curiosity landing!! http://t.co/b4sk0phy #msl]
[Rank 29] : Tweet: [g+: virtual landing party for curiosity google+ hangout #msl http://t.co/av6y9exp]
[Rank 30] : Tweet: [75 minutes to show time. i will be hosting a curiosity landing hangout info @ http://t.co/dcf3zebw]
[Rank 31] : Tweet: [cool time to watch the curiosity landing]
[Rank 32] : Tweet: [nice google+ hangout for mars curiosity here http://t.co/ggzcdhw0]
[Rank 33] : Tweet: [screenshots of @hicommander at the #msl hangout now available!]
[Rank 34] : Tweet: [@marsroverdriver is live in our curiosity landing hangout!! http://t.co/rarda1xo #marshangout]
[Rank 35] : Tweet: [what time is this mars landing?]
[Rank 36] : Tweet: [watch @marscuriosity on this g+ hangout. http://t.co/i7q2oum1]
[Rank 37] : Tweet: [#msl hangout is now live on the web at http://t.co/g8kdar2x #msl #csirotweetup]
[Rank 38] : Tweet: [mars rover curiosity hangout on google+. http://t.co/8u2jdrbu]
[Rank 39] : Tweet: [the hangout is on https://t.co/mjabqa4v #curiosity]
[Rank 40] : Tweet: [@rainnwilson is that 2 hours earth time or mars time?]
[Rank 41] : Tweet: [watching the google plus hangout on the #msl and watching the live animation of the space craft heading to mars]
[Rank 42] : Tweet: [google+ hangout covering the full curiosity landing!! http://t.co/lhkdvyre #msl #curiosity]
[Rank 43] : Tweet: [@badastronomer 's google+ hangout, nasa live tv, and nasa jpl live streaming video. i got this shit covered. #msl]
[Rank 44] : Tweet: [. @badastronomer or bigbrother? g+ hangout is the place for #curiosity talk.]
[Rank 45] : Tweet: [join in on the #msl #curiosity landing on the google+ hangout on https://t.co/j2ete9oe]
[Rank 46] : Tweet: [mars landing live]
[Rank 47] : Tweet: [we're having a great time on google+ hangout! @badastronomer @starstryder @universetoday @davemosher @astvintagespace #msl #marshangout]
[Rank 48] : Tweet: [rt @astro_flow: . @jpl with @iamwill to hangout with the tweeps. #msl http://t.co/sqkvfjdv]

[Rank 49] : Tweet: [rt @astro_flow: . @jpl with @iamwill to hangout with the tweeps. #msl http://t.co/sqkvfjdv]

**********Enter Query to search********

awesome view of its landing

********** RESULT **********

[Rank 0] : Tweet: [curiosity landing...awesome!]
[Rank 1] : Tweet: [@nasa @marscuriosity awesome... completely awesome.]
[Rank 2] : Tweet: [mars rover #awesome]
[Rank 3] : Tweet: [science! mars landing! awesome!]
[Rank 4] : Tweet: [#curiosity #nasa tv is awesome]
[Rank 5] : Tweet: [t-minus 1:40... awesome awesome awesome awesome awesome awesome http://t.co/nhqk4v6k #curiosity #jpl]
[Rank 6] : Tweet: [watching the landing of the mars rover on xbox. #awesome]
[Rank 7] : Tweet: [waiting for live coverage of curiosity landing on mars... awesome]
[Rank 8] : Tweet: [@washingtonpost @marscuriosity awesome!!!!]
[Rank 9] : Tweet: [anyone else watching the live coverage for the mars landing? because its gonna be awesome. #nasa #mars landing]
[Rank 10] : Tweet: [curiosity lands on mars today #awesome]
[Rank 11] : Tweet: [watching @marscuriosity and it's awesome!]
[Rank 12] : Tweet: [mars curiosity is close to landing, so awesome]
[Rank 13] : Tweet: [watching the mars landing on xbox live. this is awesome. #curiosity]
[Rank 14] : Tweet: [it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 15] : Tweet: [watching the mars lander #nasa #awesome]
[Rank 16] : Tweet: [curiosity is landing on mars soon. this is awesome..]
[Rank 17] : Tweet: [curiosity is now on its own..]
[Rank 18] : Tweet: [watching the mars rover landing. this is awesome stuff.]
[Rank 19] : Tweet: [watching the #ustream by #nasa for the #curiosity landing

#awesome]
[Rank 20] : Tweet: [@marscuriosity whats the view like?]
[Rank 21] : Tweet: [its all about mars!!!]
[Rank 22] : Tweet: [i'm so all over this mars landing. awesome!]
[Rank 23] : Tweet: [this is so exciting and awesome! #curiosity]
[Rank 24] : Tweet: [curiosity rover is now on its own...]
[Rank 25] : Tweet: [i can't wait for #marscuriosity to start its street view work for google.]
[Rank 26] : Tweet: [the way they control curiosity is awesome.]
[Rank 27] : Tweet: [curiosity rover is 1 hour away from landing on mars! #awesome! :d]
[Rank 28] : Tweet: [mars is about 12° in the view from the spacecraft]
[Rank 29] : Tweet: [watching this mars curiosity landing live on xbox right now...awesome]
[Rank 30] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 31] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]

[Rank 32] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 33] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 34] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 35] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 36] : Tweet: [rt @thekevindent: it would be awesome of curiosity landed on a cat! unlikely but awesome!]
[Rank 37] : Tweet: [go curiosity! you can do it!
#curiosity #awesome
#mars]
[Rank 38] : Tweet: [@marscuriosity good luck! have an awesome time on mars!]
[Rank 39] : Tweet: [less then 1 min mars landing awesome]
[Rank 40] : Tweet: [“@marscuriosity: i'm inside the orbit of deimos and completely on my own. wish me luck! #msl” / awesome]
[Rank 41] : Tweet: [#marslanding. this is going to be awesome! #curiosity]
[Rank 42] : Tweet: [@marscuriosity so awesome!!!!!! so exciting!!!!]
[Rank 43] : Tweet: [watching the @nasa mars rover lander! pretty awesome!]
[Rank 44] : Tweet: [also, to commemorate how awesome curiosity is, here's an awesome comic from the awesome @zachweiner http://t.co/xhvdkukc]
[Rank 45] : Tweet: [#jpl is so awesome!! super excited for the msl landing!!]
[Rank 46] : Tweet: [@marscuriosity good luck! you must have a great view?]
[Rank 47] : Tweet: [the @marscuriosity has its own twitter!! how awesome is that!! touching down on mars soon now! #nerdalert]
[Rank 48] : Tweet: [its going down on #mars !!!]
[Rank 49] : Tweet: [its awesome "@wired: watch live: nasa's curiosity rover attempts to land on mars in under 2 hours. http://t.co/qcet1k7i"]

Part2: PageRank

Implemented the classic PageRank algorithm on tweet corpus. Rather than scoring each tweet, we calculate the PageRank score of users. We build our graph structure based on @mentions.
If a user is never mentioned and does not mention anyone, their pagerank is zero. Assuming all nodes start out with equal probability and the probability of the random surfer teleporting is 0.1.
==========

Precision : .00001
Number of Iterations to converge : 42

1 : marscuriosity ==========>> 0.033434
2 : nasa ==========>> 0.007620
3 : iamwill ==========>> 0.003481
4 : davelavery ==========>> 0.003280
5 : sethmacfarlane ==========>> 0.001866

6 : nasa_espanol ==========>> 0.001395
7 : badastronomer ==========>> 0.001307
8 : nasahqphoto ==========>> 0.001261
9 : 1catfishknight1 ==========>> 0.001149
10 : sekerekgerg ==========>> 0.001145
11 : nasajpl ==========>> 0.001064
12 : jenna_marbles ==========>> 0.000862
13 : rainnwilson ==========>> 0.000847
14 : comedyposts ==========>> 0.000832
15 : kelly_heather ==========>> 0.000678
16 : msl_101 ==========>> 0.000651
17 : astro_flow ==========>> 0.000623
18 : gselevator ==========>> 0.000560
19 : 24horastvn ==========>> 0.000510
20 : starstryder ==========>> 0.000510
21 : marsroverdriver ==========>> 0.000482
22 : jovemnerd ==========>> 0.000472
23 : astro_sal ==========>> 0.000449
24 : fellipec ==========>> 0.000444
25 : pepitoch ==========>> 0.000415
26 : michaelianblack ==========>> 0.000412
27 : youtube ==========>> 0.000391
28 : boingboing ==========>> 0.000390
29 : scishow ==========>> 0.000387
30 : buzzfeedandrew ==========>> 0.000359
31 : urbanastronyc ==========>> 0.000358
32 : phirm ==========>> 0.000357
33 : rmaza2008 ==========>> 0.000356
34 : catherineq ==========>> 0.000356
35 : washingtonpost ==========>> 0.000355
36 : universetoday ==========>> 0.000348
37 : el_universo_hoy ==========>> 0.000341
38 : retweetthesongs ==========>> 0.000335
39 : wired ==========>> 0.000334
40 : neiltyson ==========>> 0.000328
41 : mashable ==========>> 0.000326
42 : jpl ==========>> 0.000322
43 : lindseymgreen ==========>> 0.000313
44 : carsonmyers ==========>> 0.000309
45 : rossneumann ==========>> 0.000302
46 : santicontreras ==========>> 0.000302
47 : claratma ==========>> 0.000295
48 : jonrussell ==========>> 0.000293
49 : newscientist ==========>> 0.000290
50 : cnnee ==========>> 0.000282

Part3: TF-IDF and PageRank

How to run: python part3.py

Developed an integrated tweet ranking system by integrating the cosine similarity (per tweet) and PageRank score (per user).
In this Part we Calculate the Query Dependent Page Rank of the users and then combine with the Cosvalue of the text results with corresponding query dependent user Rank, we calculate the final result.

Limitations of TF-IDF:

1. Large documents gets poor similarity score.
2. Keywords in query should be precisely matched with document terms else no results.
3. User with low page rank gets higher weightage for the basis of its content only.
4. If Somebody copied the content from the very good webpage, then both of them will be treated same.

Limitations of Page Rank:

1. PageRank algorithm considers the single pagerank vector created using the link structure of the web,to calculate the importance of each webpage depends on its inlink and their link structure and thus many users can take negative use of this process by creating false and spam structure.

2. This pagerank is independent of the any query result and for any query pagerank will give same result thus if some page has very good content but low rank will not be in close to top results.

Steps to Calcualate:

- First find the Cosine factor of each document with query.
- for each each user, we check the tweets he has done and find the maximum cosine value among all its tweets cos value. This calculation tells us that particular users tweets are how much similar to our query. Here we can choose average or maximum value but i have choosen maximum becasue there might be case when user with top rank has done a tweet with high cos value so it should have given top priority, but if i have choosen avg value of their tweets cosine, then top user with top tweet gets lower in rank.
- So We calculate the query dependent page rank by multiplying this max CosFactor in page rank formulae, which diminish the rank of the user if its tweets are not so similar.

  - **R(A) = (1 - d) + d * SUM ((PR(I->A)/C(I)) * CosFactor**

Where:

- PR(A) is the PageRank of your page A.
- d is the damping factor, usually set to 0,85.
- PR(I->A) is the PageRank of page I containing a link to page A.

- C(I) is the number of links off page I.
- PR(I->A)/C(I) is a PR-value page A receives from page I.
- SUM (PR(I->A)/C(I)) is the sum of all PR-values page A receives from pages with links to page A..
- CosFactor is maximum cosine value among all tweets of this user.

- Now we have got a Query Dependent Page Rank.
- Now while calculating the cosine similarity of the tweet with query we will multiply the Rank Factor of that tweet in the formulae.
- **Here we used Zipf's Law for diminish the Factor of Rank of the user.**

- So our New Cosine Value of Document whose user is U is :

  - **CosValue = old_cosValue * 1/(1+log(querydependendRank[U]))**

  - So we have considered the Part of User Rank for document results by cosine factor.
  - If document with high cosvalue but low page rank gets diminish cos value and user with low cosvalue but high pagerank gets its document to promoted in query results.

Output we get are very good in the sense that related to the query and by the top rank Users.

Output :

Query :

marscuriosity landing live on times square

========== **Query Based Rank** ==========

Rank: 1: :marscuriosity: 0.000886780942314
Rank: 2: :nasa : 0.00023714652872
Rank: 3: :nasajpl : 2.48113452553e-05
Rank: 4: :msl_101 : 2.40950309205e-05
Rank: 5: :gselevator: 2.33069623307e-05
Rank: 6: :starstryder: 1.9828150698e-05
Rank: 7: :badastronomer: 1.79974518259e-05
Rank: 8: :kelly_heather: 1.57640627371e-05
Rank: 9: :jenna_marbles: 1.55763566967e-05
Rank:10: :mashable : 1.45125526525e-05

========== **Query Based Result** ==========

********** Query Based RESULT **********

[1]:[Cosval: 0.126841113571]: Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[2]:[Cosval: 0.119432163539]: Tweet: [rt @urbanastronyc: @starstryder screen at times square w @marscuriosity on it. http://t.co/xas6gt1c]
[3]:[Cosval: 0.0978075173728]: Tweet: [if you are in times square watching msl landing can you please tweet a picture we can share on our google hangout? #msl #msledl]
[4]:[Cosval: 0.0942947282748]: Tweet: [rt @starstryder: if you are in times square watching msl landing can you please tweet a picture we can share on our google hangout? #msl ...]
[5]:[Cosval: 0.0577615219056]: Tweet: [live from mission control: @marscuriosity is 2 hours from landing. watch live on nasa tv: http://t.co/qmeferlo #msl]
[6]:[Cosval: 0.0541487419813]: Tweet: [@marscuriosity at times square!]
[7]:[Cosval: 0.0470617168369]: Tweet: [on my way to times square for the #curiosity landing.]
[8]:[Cosval: 0.040735788335]: Tweet: [watching the #curiosity landing from times square. because i can.]
[9]:[Cosval: 0.0391115664248]: Tweet: [kinda want to go to times square to watch the live broadcast of the mars landing. but that would involve going to times square.]
[10]:[Cosval: 0.038333131476]: Tweet: [anyone at the @marscuriosity viewing in times square?]

***********Enter Query to search**********

time for live hangout of msl landing

========== Query Based Rank ==========

Rank: 1: :marscuriosity: 0.00127737077726
Rank: 2: :nasa : 0.000308406063386
Rank: 3: :kelly_heather: 0.000101191117748
Rank: 4: :badastronomer: 4.42633749232e-05
Rank: 5: :jenna_marbles: 4.27914082431e-05
Rank: 6: :nasajpl : 3.73244413082e-05
Rank: 7: :gselevator: 2.65495474005e-05
Rank: 8: :iamwill : 2.63435450228e-05
Rank: 9: :msl_101 : 2.51803116132e-05
Rank:10: :comedyposts: 2.51306069612e-05

========== Query Based Result ==========

*************************** Query Based RESULT ***************************

[1]:[Cosval: 0.075226810462]: Tweet: [we have experts, videos, descriptions, stories, and just coolness on the live curiosity hangout: http://t.co/vcfbjny2]
[2]:[Cosval: 0.0711868076536]: Tweet: [rt @starstryder: if you are in times square watching msl landing can you please tweet a picture we can share on our google hangout? #msl ...]
[3]:[Cosval: 0.0686329904588]: Tweet: [i'm inside the orbit of deimos and completely on my own. wish me luck! #msl]

[4]:[Cosval: 0.0658456679226]: Tweet: [nbc time delay to show a race 5371 miles away: 9 hours. nasa time delay to show the @marscuriosity landing 154m miles away: 14 minutes]
[5]:[Cosval: 0.0595654075535]: Tweet: [live from mission control: @marscuriosity is 2 hours from landing. watch live on nasa tv: http://t.co/qmeferlo #msl]
[6]:[Cosval: 0.057177626544]: Tweet: [2 hours to mars, 16,300 miles away and closing fast. velocity = 8,900 mph. watch live: http://t.co/mjlj3uah #msl]
[7]:[Cosval: 0.0499708505528]: Tweet: [#msl: one hour until the landing of @marscuriosity on the red planet. are you watching? http://t.co/qmeferlo #msl]
[8]:[Cosval: 0.0426192532353]: Tweet: [for the full multimedia experience of #msl @marscuriosity, watch @nasa tv coverage and social media feeds at http://t.co/qmeferlo]
[9]:[Cosval: 0.0309817052615]: Tweet: [@1catfishknight1 #msl @marscuriosity landing is streamed live on nasa tv tonight at http://t.co/qmeferlo]
[10]:[Cosval: 0.0306277337078]: Tweet: [look at it one more time? #msl entry, descent & landing animation with tons of info: http://t.co/via1jtq6]

***********Enter Query to search**********

awesome view of its landing

========== Query Based Rank ==========

Rank: 1: :marscuriosity: 0.000439944362569
Rank: 2: :nasa : 0.000122583955482
Rank: 3: :nasajpl : 1.85898040746e-05
Rank: 4: :msl_101 : 1.44474843641e-05
Rank: 5: :buzzfeedandrew: 1.4444285469e-05
Rank: 6: :newscientist: 1.16260663001e-05
Rank: 7: :washingtonpost: 1.15521333867e-05
Rank: 8: :badastronomer: 1.09355528729e-05
Rank: 9: :kelly_heather: 1.09349215797e-05
Rank:10: :gselevator: 1.06902909729e-05

========== Query Based Result ==========

************************** Query Based RESULT **************************

[1]:[Cosval: 0.0690102844865]: Tweet: [for a view of #msl mission control without the commentary, watch here: http://t.co/wzaxvtjb]
[2]:[Cosval: 0.0505710614867]: Tweet: [yes, the mars rover has its own twitter account: @marscuriosity]
[3]:[Cosval: 0.0353458031826]: Tweet: [i'm inside the orbit of deimos and completely on my own. wish me luck! #msl]
[4]:[Cosval: 0.0340991799279]: Tweet: [curiosity landing...awesome!]

[5]:[Cosval: 0.0319607377599]: Tweet: [mars odyssey is making its turn to the proper orientation for #msl edl communications support now. currently into nominal loss of signal.]
[6]:[Cosval: 0.031825990187]: Tweet: [@nasa @marscuriosity awesome... completely awesome.]
[7]:[Cosval: 0.0314882031516]: Tweet: [mars rover #awesome]
[8]:[Cosval: 0.0285097037937]: Tweet: [1hr to entry interface! #msl is on its own executing the autonomous edl sequence. all we can do now is watch the time-delayed signals arrive]
[9]:[Cosval: 0.0278337028638]: Tweet: [rt @emikolawole: curiosity is on its own. @nasa confirms it has severed ties with #msl - http://t.co/izhxhemz]
[10]:[Cosval: 0.0277898050255]: Tweet: [the u.s. is the only nation to land a vehicle on mars and complete its mission: http://t.co/4qdj6caf #nerdolympics]

Part 4: Search Engine Optimization

My Page : http:/urjnaswxkfjjkn.wordpress.com

Approcaches/Techniques (White Hat Only) used to Boost my Webpage Page Rank:

- **Hosting your webpage on the domains which are trusted** is one of the big factor for page rank as these top domains like, wordpress, blogspot, .edu pages etc are trustworthy for google or bing and crawling of webpages hosted on these websites are fast. So i made my targed webpage on wordpress.com
- **I put Keywords in my Domain, URL and Title of the Webpage for effective ranking.**
- I try to create a sitemap of my wordpress because it is very helpful In crawling and boosting rank**. On page About me, i** also stated lot about urjnasw xkfjjkn as about me page also gets crawled first by search engines.
- Signed up for the google Webmaster tool and google analytics to check for anything wrong and see the statistics.
- **Posted very trending articles, videos , pictures on the page and tag them with the topic** so that if anybody searches for them, it should come in result like obama,gangnam style, harlem shake, johny football.
- **Links to some good webisted like Wikipedia, Texas A&M Univeristy etc for creating good outlinks**
- **Added +1 , likes and comments section to be used by the vistiors**

- From the data of google of **the most things searched on the google last day**, was most likely to be searched again on the following day, so i posted that content only like google glass, The Internship, meteorite exlosion etc.

- I used the technique of attracting lot of audience to my page by following:
    - **I keep on posted my contents update of webpage to facebook,twitter** so that people click on it like or share or watch and it also creates a backlink to my page.
    - I started **following lot of number of people** on wordpress blog, so that they might check my page in return that who had started following them and it really many reply back and people started my blog to follow as it has very updated content.

- **Insert meta tag** on the homepage of my blog **containing key words "urjnasw xkfjjkn**". Added the keywords in the headlines of the posts and the tagged keywords.

- **I put link of this page to my other webistes** like my TAMU Homepage,my personal website, other domains i owned. I also request my friends to put the url on **their website as a anchor text,** so that some rank of them got transferred to my page.

- **Posting videos on the webpage** definitely helps as instead of giving link to youtube, i fetched the youtube video and embedded on my page so that user **should spend maximum time on my webpage and increase the time of stay on my page.**

- Regularly update the web page with new contents to keep the content fresh.

Bonus

How to run: python bonus.py

So we here used the concept of topic sensitive pagerank to precalculate the pagerank for query for different topic category.
PageRank algorithm considers the single pagerank vector created using the link structure of the web,to calculate the importance of each webpage depends on its inlink and their link structure. This pagerank is independent of the any query result and for any query pagerank will give same result.

Hubs and authorities(HITS) algorithm produces a rank of the webpages depends on the user query but these ranks gets caluclated at run time and thus causes lot of delay in producing results. So we here precomputed the PageRank on some topic categories.

Here in this data corpus are tweets, so we used **the hashtags** of the tweets as the trending topics which clearly specifies the intent for text of that document. So tweet has a field "hashtags" under "entities" and contains the hashtags. while iterating over each document we fetched the hashtags of that document and created a list of hashtags in whole corpus and selects the top 16 as the topics for the Pagerank.

Then we iterate over the each tweet and if user of that tweet has put some hashtag in the text and that text is in out top 16 list, then we assign that user category of that hashtag.
if anysuch text that doesnot have any hashtag, we search for the keywords in text if lies in Top 16 hashtags, and if found we assign corresponding user that category. So a single user can be in multiple categories depending upon its hashtags or text in the tweet.
This idea seems logical to us as if user has hashtag then he must have already put its topic and it not we can infer it from its text.

So we have adjancy list for each hashtags that has users that are tagged in that category. So we modified pagerank alorithm for each topic by just assigning the inital rank value of 1/len(users of that hashtags) and for others rank to be zero. Thus we done the same for each topic and found the rank for each category.
Since we have here whole corpus about these topics so Top users we get in each category of the topics are supposed to be same and they vary for the lower ranks only but we run our algorithm on the general tweet corpus , then we will definitely get better results.

Here you can found the output as top 10 users in the top 16 topics below.

Few Changes you can observe in Bold in the Rank lists.

Topics :

[u'jpl', u'curiosity', u'marshangout', u'msl', u'space', u'nasasocial', u'mars', u'science', u'nasatv', u'fb', u'vamosmsl',
u'marslanding', u'nasa', u'olympics', u'marscuriosity', u'marte']

Output of Top 10 Users in each Topic:

******************Topic is science***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

******************Topic is jpl***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill

Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg
**********Topic is olympics**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is marslanding**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :sethmacfarlane
Rank: 4: :rainnwilson
Rank: 5: :comedyposts
Rank: 6: :jenna_marbles
Rank: 7: :kelly_heather
Rank: 8: :badastronomer
Rank: 9: :gselevator
Rank:10: :michaelianblack

**********Topic is nasa**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

*****************Topic is curiosity***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

*****************Topic is marscuriosity***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :sethmacfarlane
Rank: 4: :jenna_marbles
Rank: 5: :rainnwilson
Rank: 6: :comedyposts
Rank: 7: :badastronomer
Rank: 8: :kelly_heather
Rank: 9: :gselevator
Rank:10: :michaelianblack

*****************Topic is nasatv***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

*****************Topic is mars***********************

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill

Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is space**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is marte**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is nasasocial**********

Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1

Rank:10: :sekerekgerg

**********Topic is fb**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is msl**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is marshangout**********

Rank: 1: :marscuriosity
Rank: 2: :nasa
Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

**********Topic is vamosmsl**********

Rank: 1: :marscuriosity
Rank: 2: :nasa

Rank: 3: :iamwill
Rank: 4: :davelavery
Rank: 5: :sethmacfarlane
Rank: 6: :nasa_espanol
Rank: 7: :badastronomer
Rank: 8: :nasahqphoto
Rank: 9: :1catfishknight1
Rank:10: :sekerekgerg

References:

- Discussed the approach with Atish Patra, Amit Kumar Singh, Abhishek Jain, Amruth Kumar Juturu. (students of CSCE 670)
- Topic Sensitive Page Rank by Taher H. Haveliwala, Stanford University
- 3.NewPR-Combining TFIDF with Pagerank, Hao-ming Wang1,2, Martin Rajman2, Ye Guo3, and Bo-qin Feng1
- 4.Topic-Specific Scoring of Documents for Relevant Retrieval, Wray Buntine, Jaakko L□fstr□m, Sami Perttu and Kimmo Valtonen