Text Technologies for Data Science Assignment 4

s1140740

November 17, 2014

PageRank Algorithm

People are represented as nodes with a directed edge from a person sending an email to a person receiving the email. The weight (w) of this edge represents the number of emails the person sent. PageRanks (PR) are initialised to $1/N$ where N is the total number of people. 10 iterations of the algorithm are then executed. Each iteration S, the sum of PageRanks of sink nodes, is computed and people's PageRanks are updated using the PageRank formula.[1]

After running the algorithm for 10 iterations, klay@enron.com turned out to be the email address with the highest PageRank of 0.008027. Moreover, the email address in 6th place is kenneth.lay@enron.com which is probably the email address of the same person, CEO of Enron, Kenneth Lay. We can thus conclude that the PageRank algorithm did quite well in figuring out who the influential people are. The top 5 PageRanks can be found in table 1.

[1]
$$PR(x) = \frac{1 - \lambda + \lambda S}{N} + \lambda \sum_{y \to x} \frac{w \cdot PR(y)}{out(y)}$$

Email	PageRank
klay@enron.com	0.008027
jeff.skilling@enron.com	0.003019
sara.shackleton@enron.com	0.002961
tana.jones@enron.com	0.002855
mark.taylor@enron.com	0.002753

Table 1: Top 5 PageRanks

HITS Algorithm

Same representation as for PageRank is used. However, hub and authority values are initialised to $1/\sqrt{N}$. Each iteration the hub value of a node is updated using authority values of nodes it is pointing to[2] and the authority value is updated using hub values pointing to it.[3] Both hub and authority values are then normalised.[4]

After running the algorithm for 20 iterations, pete.davis@enron.com turned out to be the email address with the highest hub score of

[2]
$$H(x) = \sum_{y \leftarrow x} A(y)$$

[3]
$$A(x) = \sum_{y \to x} H(y)$$

[4]
$$\sum_{x} H(x)^2 = 1 = \sum_{x} A(x)^2$$

0.999281. This is consistent with his role description in `roles.txt` - broadcast proxy for auto-generated emails. Other than that there were no abnormalities. Email addresses with top 5 hub scores and top 5 authority scores are in tables 2 and 3, respectively.

Visualizing key connections

I visualised connections between 10 people with the highest PageRank. It seemed that the algorithm ranked people with great influence highly and thus visualising their connections could provide some additional information about the scandal. I used the networkx library[5] to create a graph of these connections which is then outputted into a file `graph.dot` and visualised with graphviz.[6]

I used information from `roles.txt` to assign names and roles to email accounts of Enron employees. Furthermore, the number of emails exchanged between two people is displayed at the tail of an edge in red colour.

My algorithm does not require any additional manual tuning which means that it is very automatic but that it is also quite simple.

Email	Hub score
pete.davis	0.999281
bill.williams	0.032970
rhonda.denton	0.010408
l..denton	0.006774
grace.rodriguez	0.005825

Table 2: Top 5 Hub scores

Email	Auth score
ryan.slinger	0.384187
albert.meyers	0.384177
mark.guzman	0.383849
geir.solberg	0.383764
craig.dean	0.355581

Table 3: Top 5 Auth scores

[5] NetworkX, https://networkx.github.io
[6] graphviz, http://www.graphviz.org