#-grams: Twitter Pulse from Hashtag Co-Occurrence Networks

Darshan Santani, Idiap Research Institute and EPFL
Daniel Gatica-Perez, Idiap Research Institute and EPFL

1 Purpose of the Visualization

The objective of the visualization is to understand and explore the hashtag co-occurrence dynamics on four *eventful* days in November 2012. The visualization highlights the utility of Twitter as a medium to capture global and local events via event-specific ephemeral hashtags, while at the same time the visualization tracks the ubiquity of non-ephemeral hashtags (e.g., news, games, music, etc.) on each selected day. Various key observations stand out, two of them are listed below:

1. It is interesting to observe how Twitter's "follow" jargon in different languages are linked with each other – #teamfollowback is connected to #sougofollow via #siguemeytesigo (#siguemeytesigo and #sougofollow means reciprocal follow in Spanish and Japanese respectively). This trend is observable across all days under study.

   One of the advantages of visualizing the co-occurrence of hashtags is that the resultant network adds semantic context to help decipher the meaning of hashtags which otherwise seem quite unintuitive, like the ones in this example. If the semantics of a hashtag in the sub-cluster is known, it most likely provides the meaning for other connected hashtags.

2. Another interesting observation is spotted on 14 Nov. 2012, the day when the workers' union in Spain called for a nation wide strike. The hashtag co-occurrence cluster centered around #14n emerged but surprisingly the resultant sub-cluster was not part of the largest connected component (LCC) on that day (though the cluster got reduced and joined the LCC the next day). Similar trend is observed for the sub-cluster around #peopleschoice hashtag on the Thanksgiving day (22 Nov.). Apparently earlier in the week that year, the nominees for the People's Choice Awards were announced in the US.

Rather than exhaustively listing all the observations, we ask users to explore the visualization and come up with their own interpretations and findings.

2 Dataset

We have used the publicly available Twitter data which is released as part of this challenge [1]

3 Description

The hashtag co-occurrence network is an unweighted, undirected graph, where each node is a hashtag, and an edge exists if two hashtags co-occur in a tweet at least 10 times. We have selected only those tweets in our corpus which contain exactly two hashtags. Nodes (and their first order neighbors) are highlighted and colored based on micro-topics. We have chosen four

non-ephemeral topics – news, twitter follow jargon, games, and music – which are prevalent on each day. In addition, we have also chosen an event-specific micro-topic for each of the four days. A micro-topic is chosen based on a naive string matching, i.e., for "news" micro-topic we have selected all those hashtags which contain "news" as a substring e.g., #news, #breakingnews, #technews, #newsjp, etc.

Nodes belonging to multiple topics are highlighted using a different color and are sized twice the size of a typical node, to signify their mixed membership; but otherwise all the nodes are equally sized. To determine the layout of nodes while drawing the network, we have used the force-directed Fruchterman Reingold layout algorithm [2].

The visualization is conceptualized and created in R! programming language [3], while the design has been finalized using Inkscape under the Linux environment.

4 Miscellaneous Information

The visualization is best viewed in Adobe Reader, which allows you to zoom-in at higher magnification to explore the intricacies of the network. On the Linux platform, we have tested the zoom-in capabilities of other PDF reader tools like Evince and Okular, but these tools currently do not support zoom-in beyond 400%, and that resolution is not sufficient to clearly view the labels on the nodes.

References

[1] K. McKelvey and F. Menczer, "Truthy: Enabling the Study of Online Social Networks," in *Proc. 16th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW)*, 2013.

[2] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*. 2014.