How to produce a quick summary of top tweets using NodeXL (new and updated article)…

…and more advanced analysis in a search for a complete extract.

Twitter can be an excellent way to share and source information, but the volume of information can quickly become overwhelming. When there are hundreds or thousands of tweets about a topic in a short period it becomes inefficient to rely on browsing and searching. Twitter's own search algorithms are not always very reliable at identifying top content. Also, Twitter appears to serve up recommendations based on your previous activities, so there is a risk of becoming stuck in a bubble and not seeing the fuller picture, for example around a health awareness campaign.

There are, however, relatively simple ways (eg NodeXL) to extract a full record of tweeting around a health campaign or conference, which can then be summarised to produce a "top tweet" record of the topic of interest. This can save a lot of time: it can help identify the most shared content (positive and negative) and also reveals the top tweeters (people who you may wish to follow or ask to support your work). While NodeXL reports produce useful summary reports, which include measures of influence and connectness, there can be potential quirks, pitfalls and biases if we rely on these measures alone, so it is always best to look at the individual tweets too, as described here. This also has the useful by-product of allowing us to summarise top tweets to share more widely.

This article updates a previous "how to" blog about producing a "top tweet" analysis from a NodeXL extract.

The article is in two parts:

1) how to produce a "top tweet" summary from a NodeXL extract

2) how to produce a NodeXL extract when the campaign lasts more than a few days, or when initial attempts at extracting data are unsuccessful.

Part 1) Producing a "top tweet" summary from a NodeXL extract

I have been refining my methods over recent months, and hopefully this updated approach will be quick and easy to use. Please contact me via Twitter (@gmacscotland) if you have questions. This is a first draft and will be refined over time.

I have produced this as a PDF file rather than a "blog" because it is less likely to corrupt key text required to refine the extract in Excel.

The tool that I used to list the tweets previously (Storify) is closing in May 2018, so I have updated that section too for an excellent replacement – Wakelet, developed by a start up in Manchester, UK. You can see a list of my Wakelet summaries (some new, some imported from Storify) here: https://wakelet.com/@scotpublichealth.

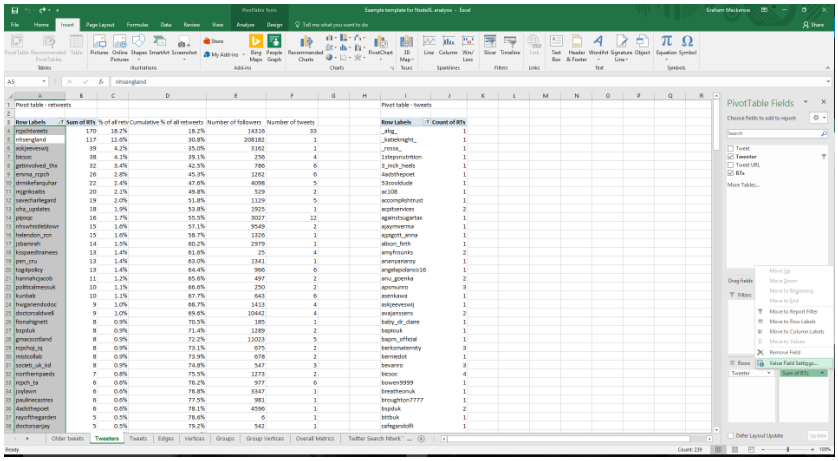
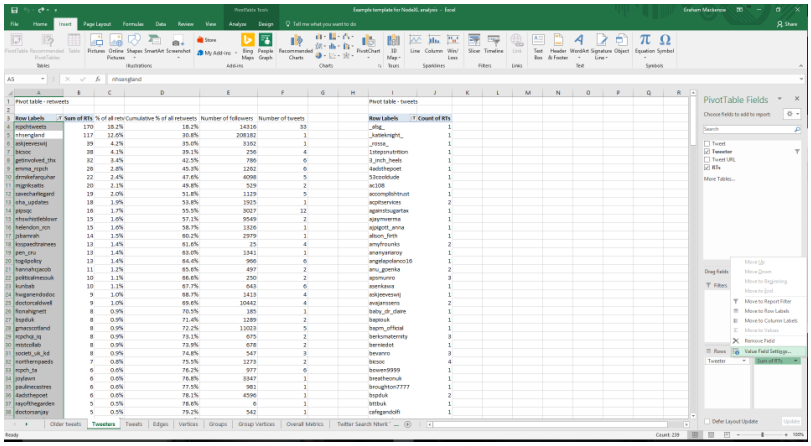
Once you have tried the approach out with a couple of NodeXL extracts you should be able to produce a "top tweet" analysis of even the largest Twitter campaign or conference in about 20 minutes. This is invaluable when summarising at the end of an event, or even when having a quick look mid campaign.

The approach is as follows. You will need some Excel skills, or some patience and trial and error to work through, understand and learn the different steps. You will need this "Example template for NodeXL analysis" (save in your analysis folder, or look in Downloads). The example uses tweets with "RCPCHtweets" (tweets by or about the Royal College of Paediatrics and Child Health).

This template file shows you the end result of an analysis in the "tweets" and "tweeters" tabs. It shows some of the formulae you can use to count tweets, tweeters, search for hashtags and more. Click into some of the cells to look at the workings. You can find further explanation in this Powerpoint file. If you're using a PC then you will need to make use of the left mouse button (to select) and right mouse button (to show you options, including "delete", "sort" in a pivot table and more).

1) Open up a copy of your "template file".	a) Minimise this, and move on to step 2.
2) Identify a NodeXL extract to analyse	a) If you have run a NodeXL analysis yourself then you should save a copy on your PC. b) Alternatively, you can download other people's data from the NodeXL Graph Gallery. Scroll down to the bottom of the webpage of the NodeXL report and click "Download the Graph Data as a NodeXL Workbook". Save to your analysis folder (or look in Downloads folder).
3) Identify the main data that you want to use (eg tweet, URL for that tweet, tweeter, number of likes and retweets, hashtags)	a) Open up the Excel file. b) Find the "Edges" tab at the far left of the workbook. c) Go to the "insert extra columns here" column, select it, and insert 8 new columns. d) Go back to your "template file" and copy the cells highlighted in red (two rows) from "Edges" and paste into the same location in your NodeXL extract. There are two rows – one that labels the columns, and the other that looks for original tweets in your extract by excluding rows starting with "RT". The cells below this second row should be automatically populated. Scroll down your sheet a little to check that this has happened. If not, then copy and paste the second row and paste all the way down to the bottom of the sheet.
4) Remove duplicates	a) Set up a new sheet within your workbook. Name this "selected tweets". b) Copy the columns with your data (click on the letters at the top of the columns to select the full columns). Paste these columns into your new sheet "selected tweets". c) Remove the top row (click on the number 1 at the far left of the top row and delete). d) Select all the data in this sheet. Order by "tweet URL". Delete the blank lines (using the same approach as you used in 4a). e) Label a new column to the right of the rest of the data as "Duplicates". Add the following text: =if(F2=F1,1,"") into cell M2 to identify duplicates. Copy this right down to the bottom of your data. f) Select this whole column (select column M). Copy and paste (using special "values").

	g) Select all the data in this sheet again, including the duplicates column, and order by duplicates (from small to large). Delete the rows with a "1" in duplicate column. You should now have a list of unique tweets. To see what this looks like, have a look at the "tweets" tab of the "template".
5) Order the tweets as appropriate for the task in hand	Think through the order that would fit best with your data. NodeXL identifies tweets that were tweeted or retweeted during the period of the extract. Accordingly, you might have some older tweets in your extract. You might want to move them, so that you're only working with tweets posted during the period of the extract. The Powerpoint file shows you a quick way to identify older tweets from your analysis, and apply other rules at the same time, but you might prefer to work manually at first to understand the data. Once you have the tweets that you want to work with, you can order them using simple rules based on information already in your sheet, marking your selected tweets. Rename the "duplicates" column as "selected tweets". Here are some examples of ways to sort: a) Top tweets for the whole period, based on combined number of retweets and likes: select all the data in the sheet. Go to "Data, sort", and select the relevant field (ie combined retweets and likes, descending order). Have a look at the top tweets to see whether they provide a suitable summary. Select each tweet that you plan to use by putting a 1 in the "selected tweets" column. b) Top tweets for each day of a conference or campaign: again, select all the data, but this time you will use two fields – day and number of retweets. Again, inspect the top tweets for each day and put a 1 in the "selected tweets" column beside your chosen posts. Whether you use a manual approach at this stage, or when you come to produce your summary (step 7) you should have a look at the content – exclude spam and offensive materials. If you have a large extract, or just want to do things quickly, you can automate the rules for selecting tweets – eg by adding the following to the "selected tweets" column and copying down the column: =if(I2>=10,1,"") . This will select tweets with at least 10 likes or retweets. You may find that the number of tweets and retweets varies by day of conference or campaign, and you can refine the rules for each day. Decide whether you want to display your tweets by popularity (eg descending order of combined retweets and likes) or chronologically (using the date field). Select all the data and sort accordingly, adding in "selected tweets" as your first sort term. You can select more than one term to use in the sort. Make sure you select all the columns you want to sort.
6) Summarise the main	a) Count the number of tweets and retweets using the "count" function. Look at the blue box in "tweets" tab in the example

statistics (eg number of tweets, retweets, tweeters, and the top tweeters)	template and click in to the results to see how “count” works. The dollar symbol used in some of the formulae makes sure you keep looking in the same cells as you move down a column or, depending how you set things up, across a row. These are called “absolute cell references”. You will need to adjust the formulae depending on the amount of information extracted. b) Count the number of tweeters as follows: select columns with tweeters and retweets and all columns in between. Select “Insert pivot table”. You can see this in the top left corner of Excel.  c) Configure the pivot table with “tweeter” as row label and “retweets” as value label. Select “value field settings” as either “count of retweets” (which gives the number of tweets that each tweeter posted), or “sum of retweets” (which gives the number of retweets). Click the pivot table and order to identify the accounts with most tweets or retweets. Use “count” to identify the number of tweeters.  d) You can use the list produced in section 6c to identify the people tweeting most, and retweeted most. You can also look at this in more detail to look at the accounts that received 80% of retweets. Typically this will only be a small proportion of overall tweeters. Compare this list with the list of top influencers in the NodeXL report in the NodeXL Graph Gallery. To see what this looks like, have a look at the “tweeters” tab of the example template file. You can see some extra columns, and the workings – eg finding the number of followers for each tweeter by looking across to the “vertices” sheet (note that vertices sheet needs to be sorted alphabetically by vertex (tweeter) for this to work).

7) Extract to a Wakelet summary	a) If you do not have a Wakelet account, set one up. b) Log in to Wakelet. c) Click “New collection/ Story”. Give it a title and description d) Set up your Excel file and browser with Wakelet side by side on the screen (click on the restore down button in the top right corner of the Excel and browser windows). e) I find it easiest to add tweets from the bottom up. Copy an individual tweet URL from your Excel sheet and paste it into the “add a link” box. Move up the list, adding your selected tweets. f) Provide context so that people know how you have produced the summary: Add the URL for your NodeXL report and this will be displayed with text and a map in the Wakelet. g) Add text (“write something”) above the NodeXL link to explain briefly what you have done. Provide the summary statistics – eg number of tweets, retweets, tweeters and main influencers. See the final results here for my analysis of the data in the example template, including a link to the NodeXL report. Publish your own Wakelet summary and set to “public” view.
8) Share your summary with the top tweeters	Share your findings as appropriate – eg by tweet. You can mention the top influencers identified in stage 7 to let them know about your work. The Wakelet will appear with an image and description in your tweet. 

More advanced tasks from the NodeXL extract:

For very large searches, where you want to achieve a balance by hour rather than just by day, you can use the Excel “hour” function to extract this information from the date column.

You can run more sophisticated analysis by adding columns to the Excel sheets. One example would be adding a column for number of followers. In order to do this, select the “vertices” tab in the original NodeXL file. You will see a list of people tweeting, retweeting or mentioned in the tweets. Arrange this sheet in alphabetical order by Twitter handle (username). If it’s a very big search then you will need to copy and paste the columns with handle and number of followers into a separate sheet and use this instead. Once you have a list in alphabetical order you can use

the VLOOKUP function in Excel to look up Twitter handle and lookup number of followers. This could be useful, for example, if you wanted to stratify your analysis into people with small, medium and large numbers of followers. There's an example of using VLOOKUP in the Excel example template.

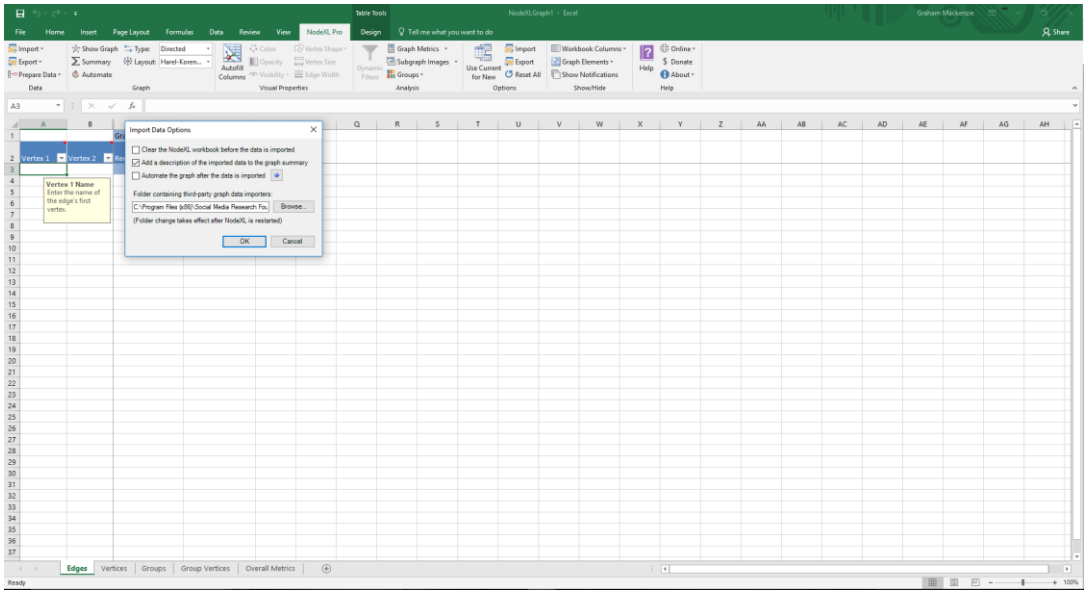
You can use the information contained in the "hashtags" column in a number of different ways. You can search out specific hashtags using the "search" function. Or you can count the number of hashtags using the "substitute" and "length" functions, looking for spaces. See https://exceljet.net/formula/count-specific-characters-in-a-cell for more.

2) How to produce a NodeXL extract for longer campaigns, or when the extracts prove elusive

This section looks at more advanced techniques in producing NodeXL extracts. You will need a copy of NodeXL (I use a licensed copy, which costs $200 per annum from Social Media Research Foundation; I have not attempted the methods described below in the free version).

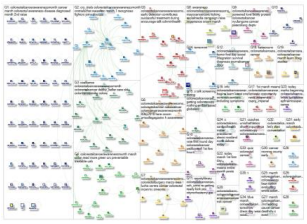
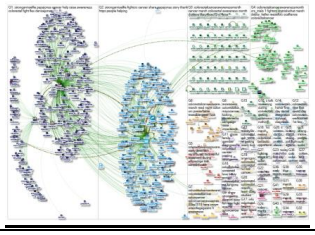
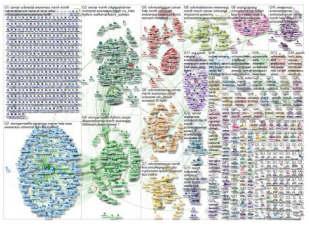
Typically NodeXL will produce 9-10 day extracts (it cannot look back further, but it will record tweets retweeted during the period of the extract). You can extract information about the top tweets as described in section 1. Sometimes, however, you will need to look beyond the 9-10 period, planning a longer campaign into the future. Other times you will discover that NodeXL is only managing to extract a much shorter period of tweets (typically because Twitter is busy, or there is a lot of activity around your topic of interest). In both cases you will need to produce more than one NodeXL extract and combine these. (There is an alternative – importing directly into an existing NodeXL extract, but it is more difficult to check the reliability of the extract, so I prefer to use the following method).

Both the above scenarios rely on combining NodeXL extracts. This is relatively straightforward, though can take time for larger extracts. You need to download the NodeXL data files in question from the NodeXL Graph Gallery site. Select the page for the report you're interested in and scroll to the bottom of that page to find the "Download the Graph Data as GraphML" link. Save this in a folder. Repeat for each NodeXL report of interest. Open a new NodeXL file. Go to import options and uncheck the "clear the NodeXL workbook" option.



Import each of the GraphML files. Run the analysis as usual (graph metrics, etc). Upload the files to the NodeXL Graph Gallery website.

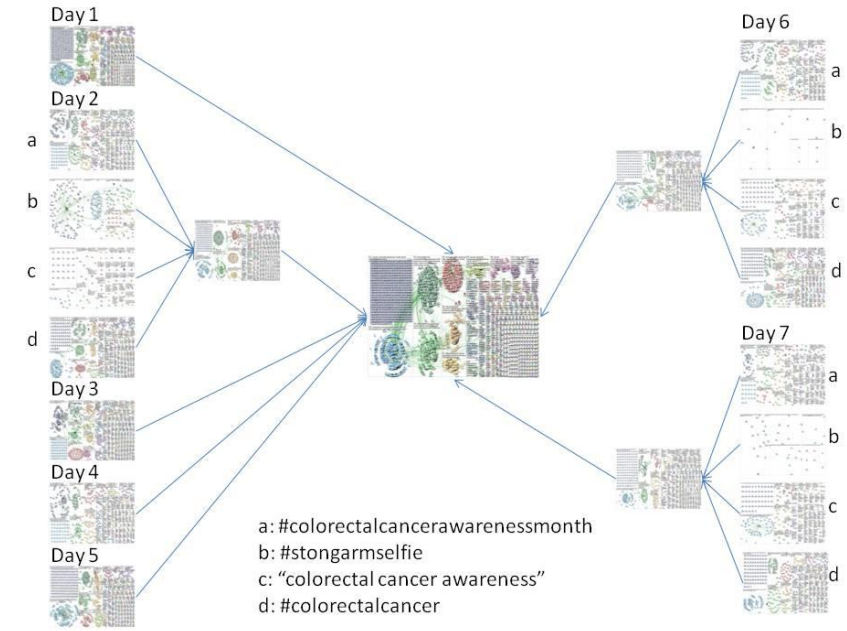
A recent example of this type of combined analysis is illustrated by week 1 of “colorectal cancer awareness month” (1-7 March 2018). I had identified a number of terms of interest related to this campaign. The sequence of graphs below shows the impact of adding in these additional terms. Think of this like a literature review – you need to look as broadly as possible to capture a complete extract, before focusing in on the items of particular interest.

1) #Colorectalcancerawarenessmonth 21 Feb to 1 March 	2) #Colorectalcancerawarenessmonth OR #strongarmselfie 20 Feb to 1 March 
3) #colorectalcancerawarenessmonth OR #strongarmselfie OR "colorectal cancer awareness" OR #colorectalcancer 20 Feb to 1 March 	

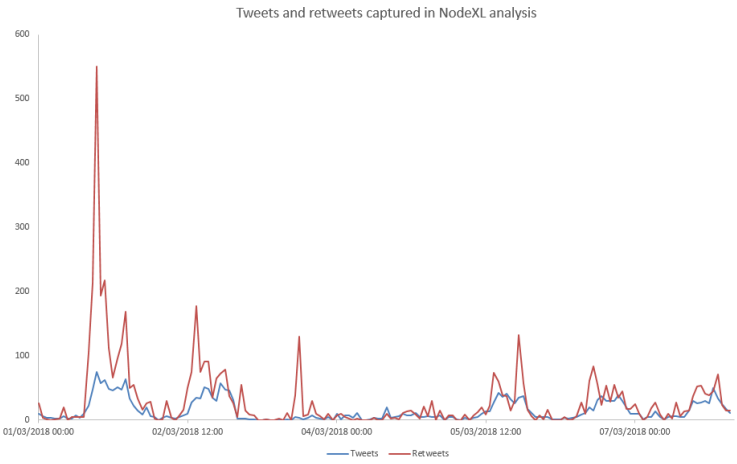
1) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=142490
2) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=142491
3) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=142493

However, attempts to produce a complete extract of the first week, in a single go, were unsuccessful. This was frustrating as in theory this extract should be well below the Twitter/ NodeXL limit of 18,000 tweets.

I therefore had to run separate extracts, initially by day, and then for individual terms for each day. I combined these as shown below. Most of these searches were complete, but I was not able to extract search 2c, which only gave 3 hours 42 minutes of the 24 hours (towards the end of the day).



Nonetheless, this gave a much more complete dataset than attempts to extract a full week in a single go. The [Wakelet summary] includes the NodeXL report and results of the top tweet analysis. The number of tweets and retweets included in the analysis are shown below (again this is produced from the NodeXL extract, using Excel date and time functions (to give whole hours) and, in a further work around, plotted using a scatterplot with lines).



Graham Mackenzie, Consultant in Public Health, NHS Lothian, 13 March 2018

[@gmacscotland]

Source materials for the colorectal cancer awareness month analysis.

Day 1:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143731
Day 2:
a) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143747
b) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143748
c) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143749 * incomplete - 3 hours 42 minutes
d) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143750
2 combined:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143768
Day 3:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143751
Day 4:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143752
Day 5:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143753
Day 6:
a) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143755
b) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143756
c) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143757
d) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143758
6 combined:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143769
Day 7:
a) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143760
b) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143761
c) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143762
d) http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143763
7 combined:
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143770

Final output
http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=143772

(*Complete, except for around 20 hours in search 2c).