Step-by-Step Guide to Workshop Activities

This guide documents the analysis steps demonstrated in the Social Media Analytics workshop during the QUT Digital Media Research Centre's Digital Methods pre-conference. This is a barebones document which does not discuss the challenges of gathering social media data, nor the ethical considerations required in doing so – you are expected to have considered these already. Also, the focus here is solely on working with Twitter data, though many of the methods and approaches will be able to be translated to working with data drawn from Facebook, Instagram, or other social media platforms.

This guide is a work in progress – any feedback appreciated (a.bruns@qut.edu.au).

Sourcing Data

The currently leading tool for sourcing Twitter data (tracking keywords, hashtags, or users within the limits of the open Twitter API) is the Twitter Capture and Analysis Toolkit (TCAT), developed by the University of Amsterdam's Digital Methods Initiative. This needs to be installed on a server, and set up to track and capture public Twitter content as required.

1. Data download:

- Download the following three datasets:
    a) Export all tweets from selection
    b) Export hashtag table
    c) Export mentions table

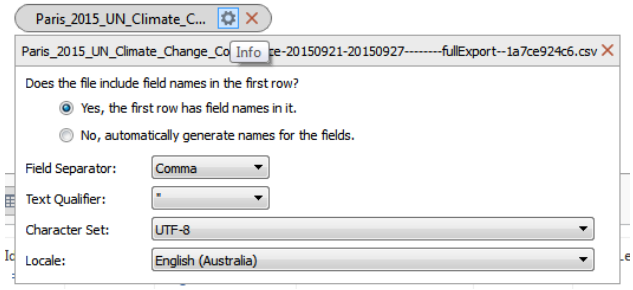
The following workshop activities draw exclusively on these three datasets.

Analysing Data

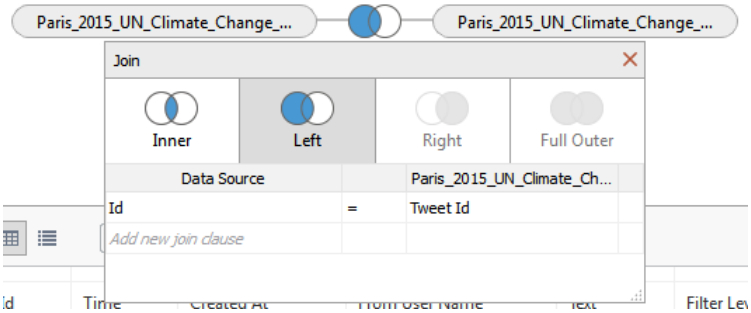
The most flexible and most powerful tool for analysing social media and other datasets at present is Tableau – a commercial software available for free to current students and under educational licences to researchers. Contrary to other standard software such as Excel, Tableau is able to work with very large datasets in a very wide range of formats.

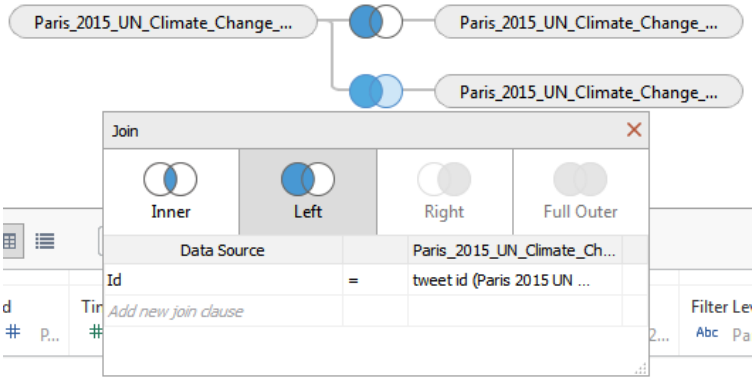
2. Loading the datasets into Tableau:

- Open Tableau, connect to data
- Choose Text file and connect to main CSV file
  (and check file format settings: comma, quotes, UTF-8, correct language for main CSV):



- Drag hashtag and mentions files into whitespace next to main file
- Change relationship to left join, and check file format settings;
  in each case, join on Id = Tweet Id:

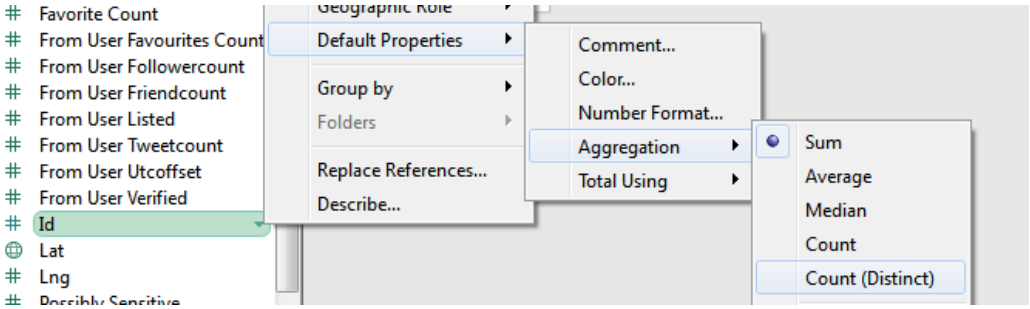




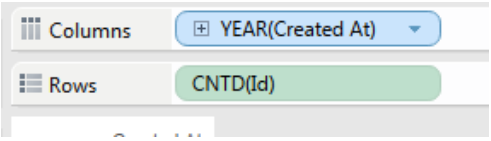
- Click Update now to inspect
- Choose Extract, click Sheet 1 to go to worksheet (this may take a moment)

3. Set and explore the basic data properties

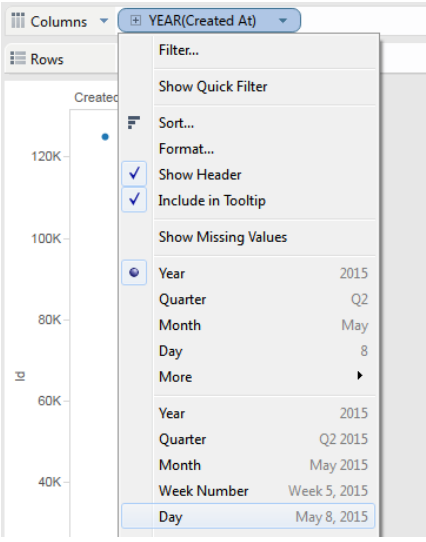
- Move Id field to Measures, change default aggregation to Count (Distinct): CNTD(Id)



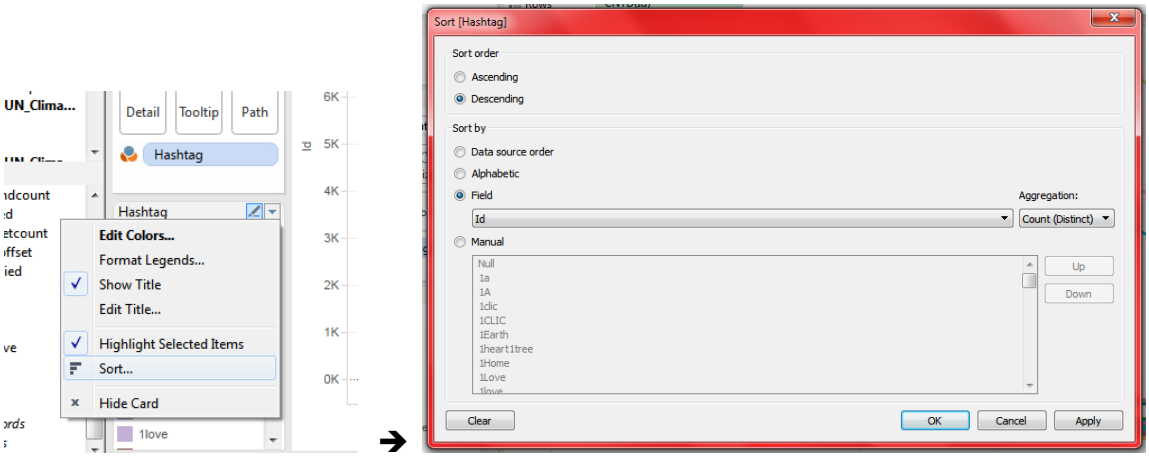
- Move Number of Records and CNTD(Id) onto Rows, and compare the figures
  a) Note the difference – Number of Records is inflated because of joined tables, so we'll always use CNTD(Id) instead
- Now move Created At on Columns, CNTD(Id) on Rows



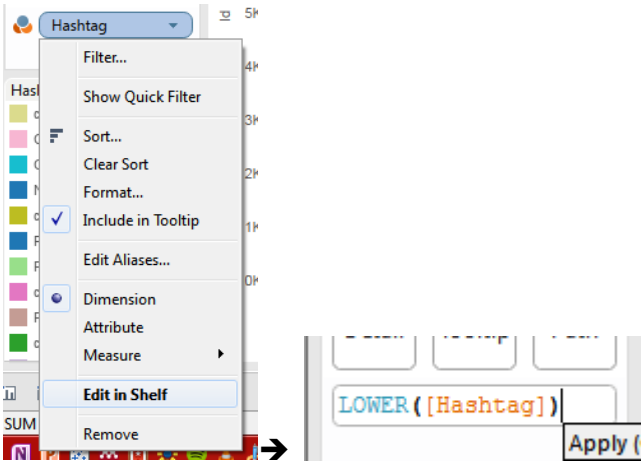
- The default scale for date fields is Year, change this to the second Day option



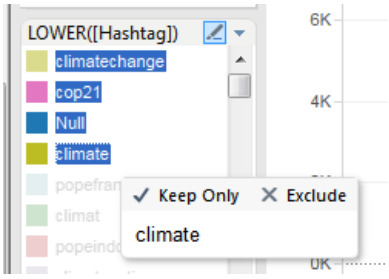
- Try Mention Type / Hashtag / From User Name / User To Name / From User Timezone / Source on Color
- Sort Color legend descending by CNTD(Id)



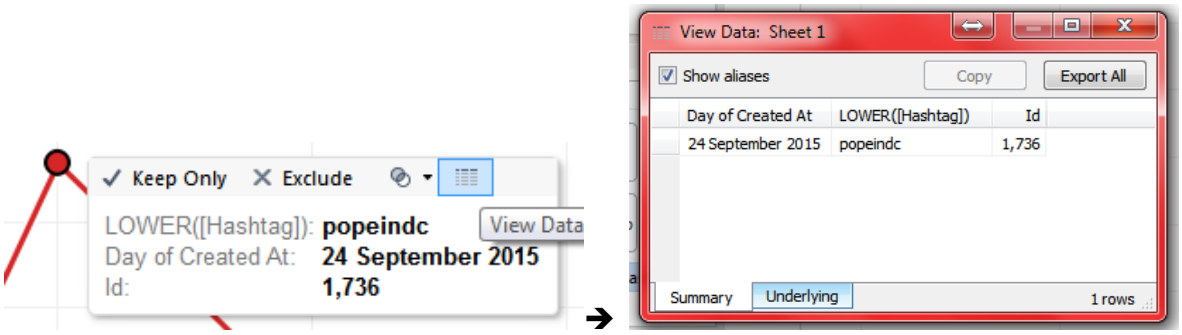
- If necessary, change case-sensitive fields to lowercase by using the formula LOWER([field]) – the easiest is to use the Edit in Shelf function for an on-the-fly conversion, e.g. for Hashtag:



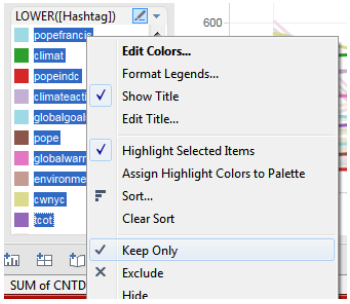
- Exclude generic terms from the list as required: select and right-click



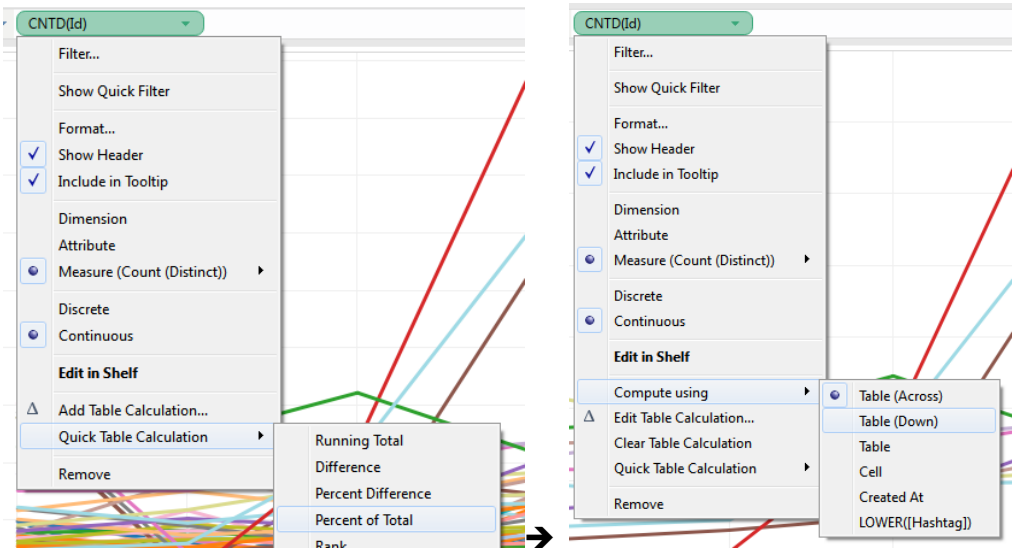
- Drill down into raw data to explore underlying patterns: select data point(s), click the data icon, and select the Underlying tab



- Limit visible items to top 10-20, using Keep Only

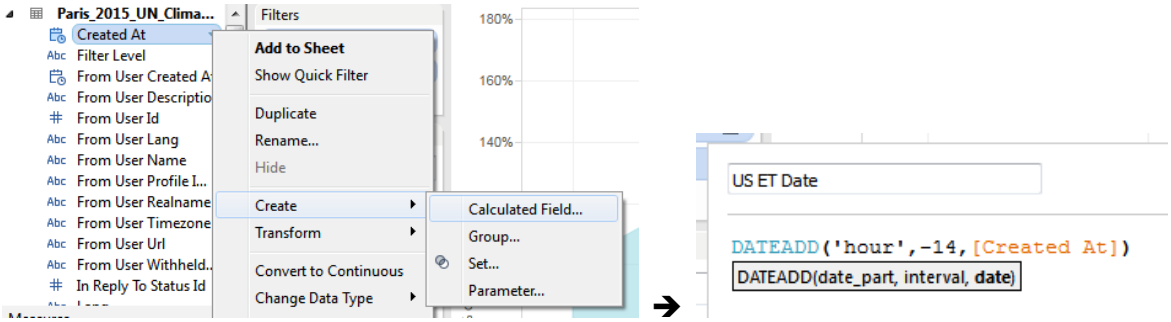


- Switch graph values for CNTD(Id) to percent of total, and calculate using Table (Down)

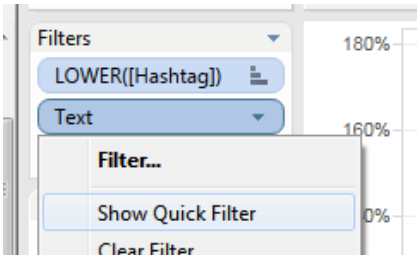


- Switch graph style to to continuous area map

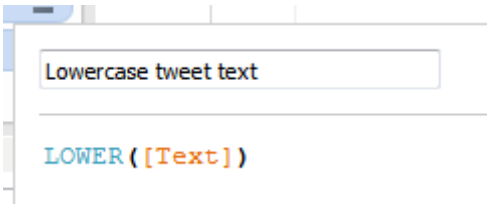
- CTRL-drag field from Color to Label
    a) Note that values above 100% are caused by multiple hashtags / mentions per tweet
- Explore switch from day to hour as unit of time
- Create localised date field to demonstrate timestamp conversion:
  e.g. DATEADD('hour',-14,[Created At]) (Brisbane time -14h = US ET)



- Create text filter as quick filter

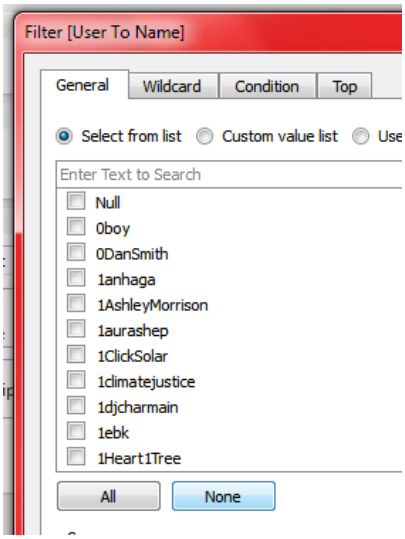


- Note case sensitivity – if necessary, create LOWER([text]) as a new field to filter on

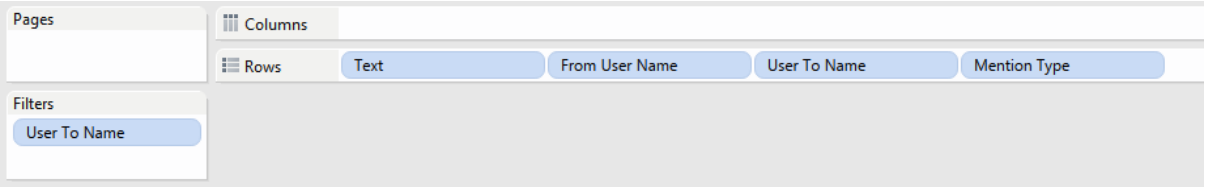


4. Exploring tweet and retweet texts, using quick filters

- Set User To Name filter, select None (NOTE: do not use To User Name)



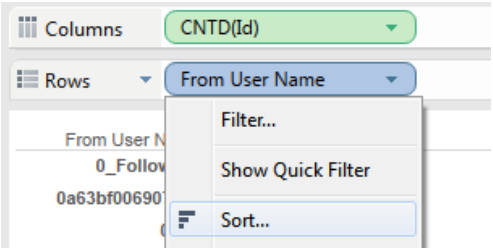
- Text, From User Name, User To Name, Mention Type into Rows – select Add all members if asked



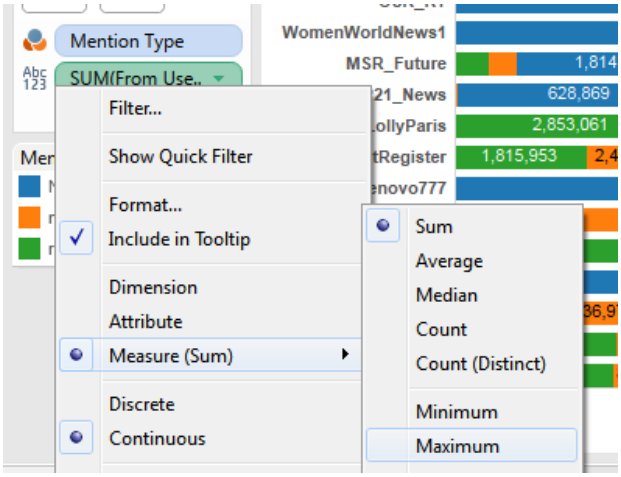
- Turn User To Name filter into quick filter
- Explore:
  a) e.g. all CNN / BBC / … accounts
  b) check (or filter) for retweets
  c) Explore other variations

5. Identifying most active / most visible users

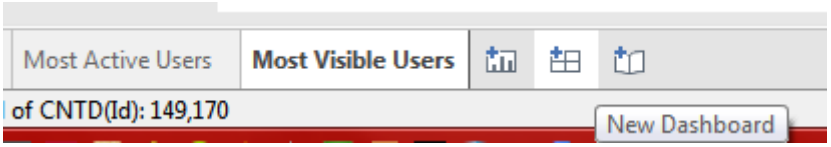
- Create two sheets:
  a) From User Name on Rows
  b) User To Name on Rows (do not use To User Name)
- CNTD(Id) on Columns
- Sort descending by CNTD(Id)



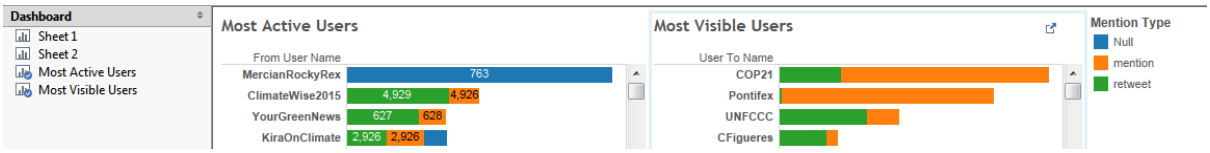
- Mention Type on Color
  - Note: inflated numbers due to multiple mention types per tweet
- In sheet a), add From User Followercount on Label – and change aggregation method to Maximum



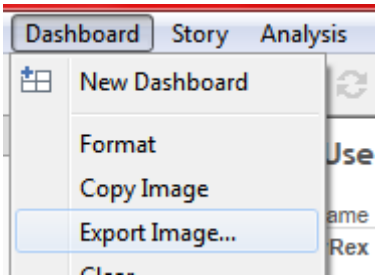
  - Note: for User To Name (most visible users), exclude Null (original tweets, not @mentioning any other user)
- Name both sheets and create a combined Dashboard



- Drag sheets side by side onto Dashboard canvas

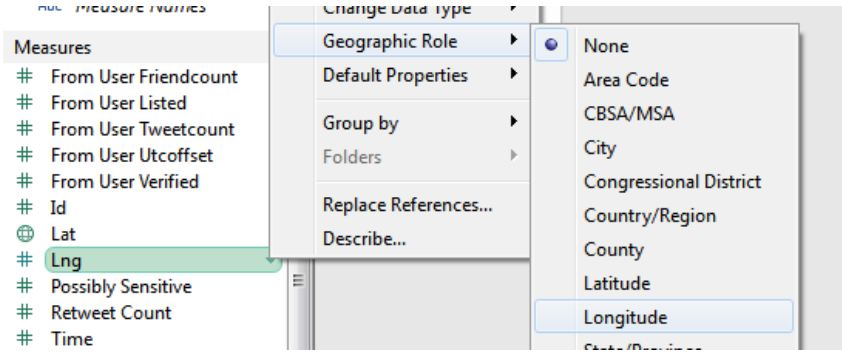


- To export to Word or Powerpoint, use copy / export functionality:

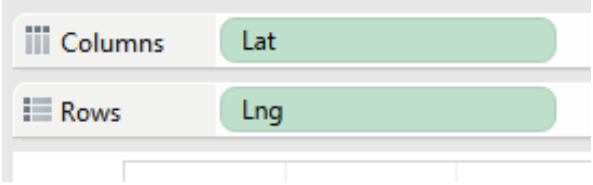


6. Geolocation

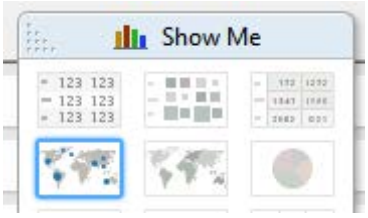
- Change Lng field to Longitude type



- Move Lat/Lng to Dimensions
- Lat on Columns, Lng on Rows



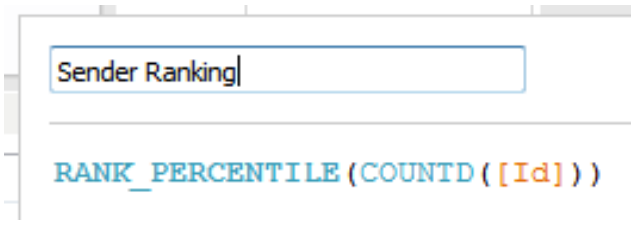
- Change to symbol map, explore



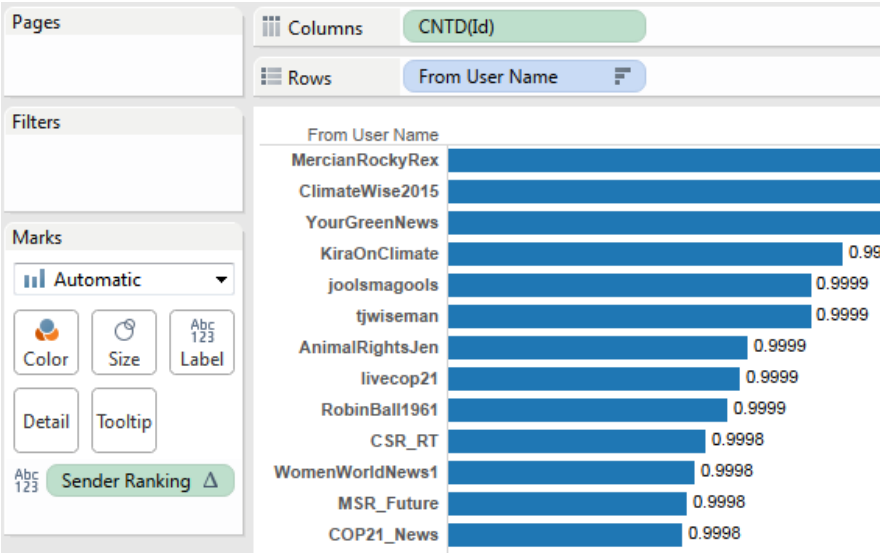
- Note: usually inconclusive due to lack of volume

7. User Percentiles

- Creating percentiles:
  - Create calculated field Sender Ranking: RANK_PERCENTILE(COUNTD([Id]))

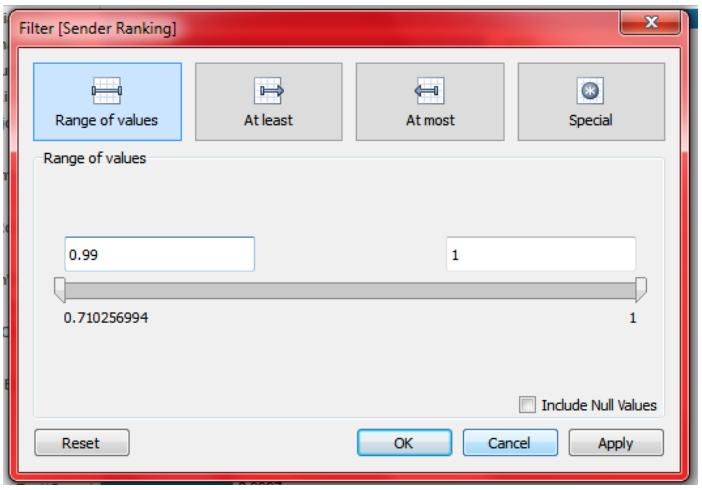


  - Graph CNTD(Id) against From User Name, Sender Ranking on Label, order descending

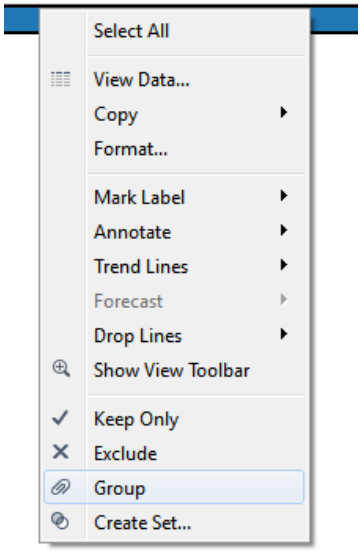


o Filter by Sender Ranking (repeat twice):

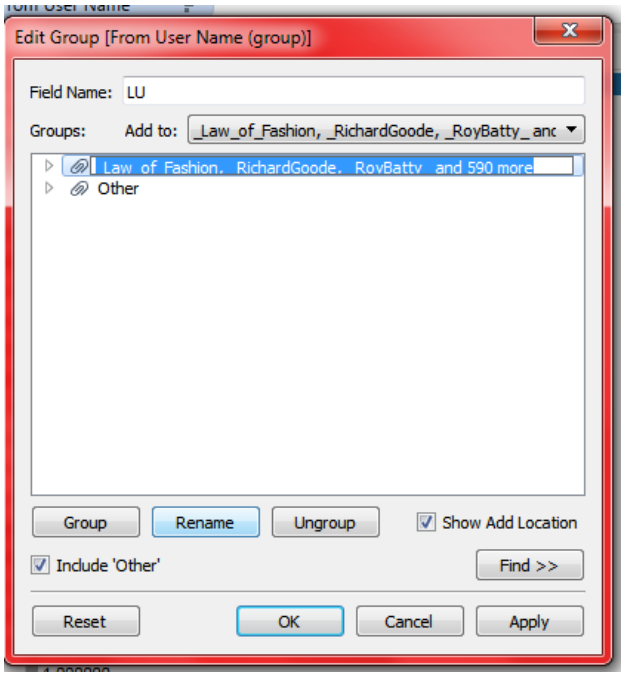
1. Sender Ranking >.99



- Select all (CTRL-A), right-click, create new Group 'LU' (lead users)



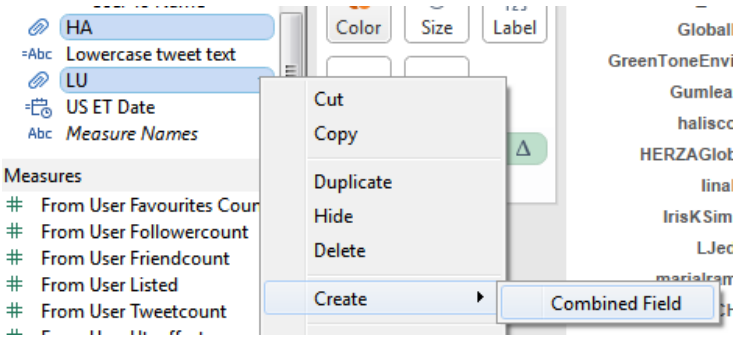
- Edit the new group:
  - change field name to LU
  - Use Rename button to rename first list to LU (lead users)



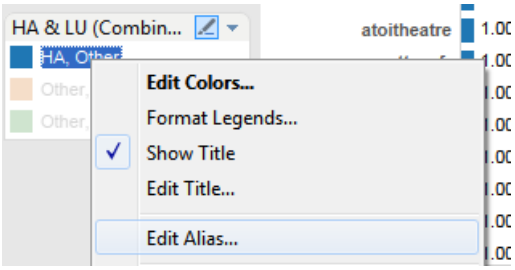
- Remove new LU group from Color

2. Change Sender Ranking filter to .90 to .99
   - Repeat previous steps to create new group HA (highly active users)

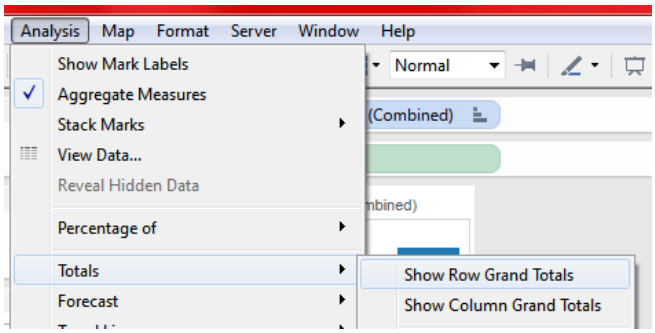
- Select LU + HA groups in Dimensions, and create a Combined Field



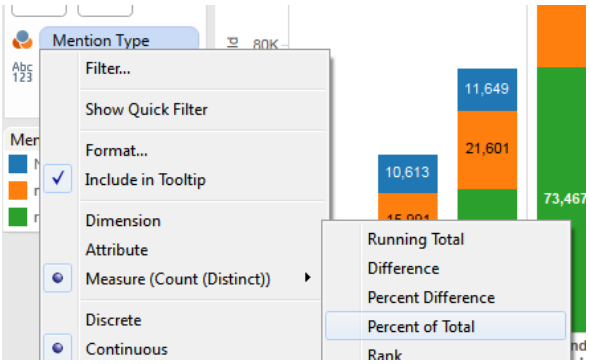
- Drag combined field onto Color to demonstrate group membership
  - Note: Sender Ranking values are now calculated separately for each category!
- Use Edit Alias to rename categories in colour legend



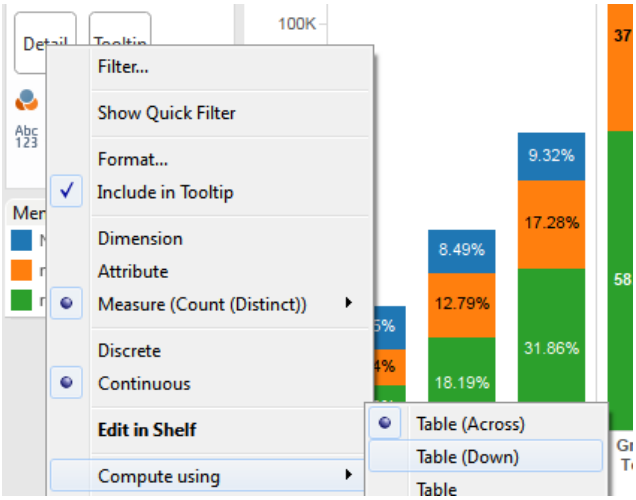
- Using percentile groups in the analysis:
  - Place combined field on Rows, CNTD(Id) on Columns
  - Use Analysis > Totals > Show Row Grand Totals



  - Mention Type on Color
  - CNTD(Id) on Label, change to Percent of Total



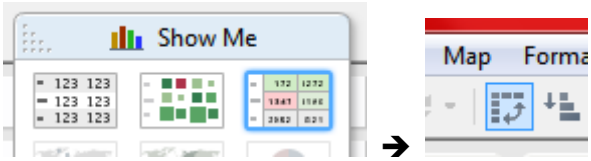
  - Compute label percentage by using Table (Down)



  - Make same changes for CNTD(Id) in Rows: Percentage of Total, Compute Using Table (Down)
  - Note: values above 100% due to some tweets being both retweets and @mentions
  - More analysis options:
    - Replace Mention Type with Source on Color, sort descending by CNTD(Id), explore

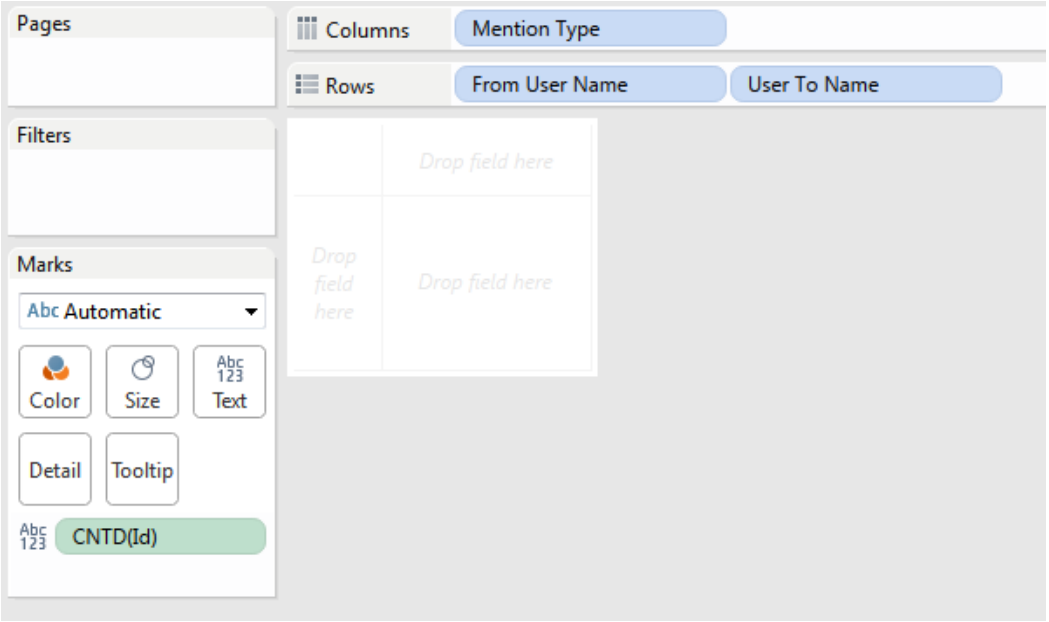
- Same with From User Verified (change to Dimension)
- Same with Hashtags
- Same with User To Name

- Examining attention to key users by percentile groups:
  - User To Name on Rows, Combined field on Columns, CNTD(Id) on Text
  - Sort User To Name by CNTD(Id)
  - Change CNTD(Id) to Percent of total
  - CTRL-drag CNTD(Id) to Color, change to red/green diverging colour scheme
  - Change to Highlight Table, and rotate table

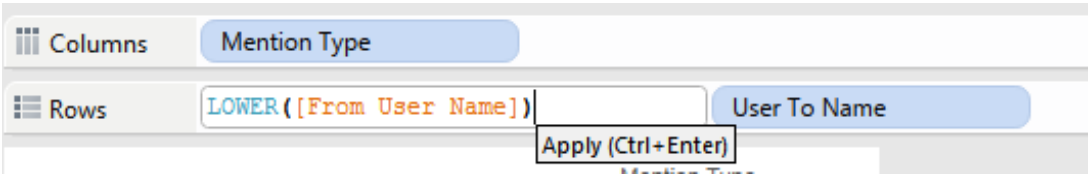


8. Exporting data from Tableau to Gephi, for network analysis

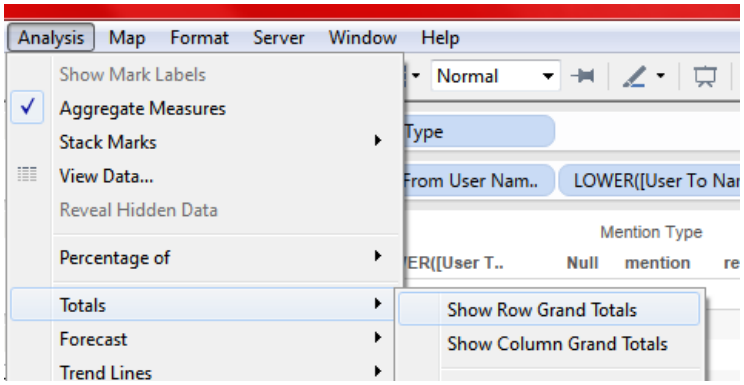
- From User Name, User To Name on Rows, Mention Type on Columns, CNTD(id) on Label



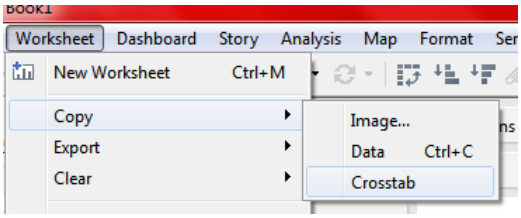
- Change From User Name / User To Name to lowercase (via Edit in Shelf)



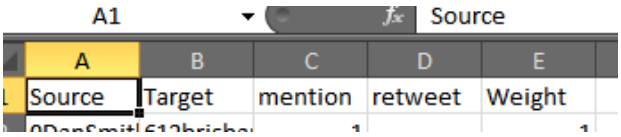
- Analysis > Totals > Show Row Grand Totals



- For large datasets, add further filters as required to limit data size (e.g. one day only)
- Worksheet > Copy > Crosstab



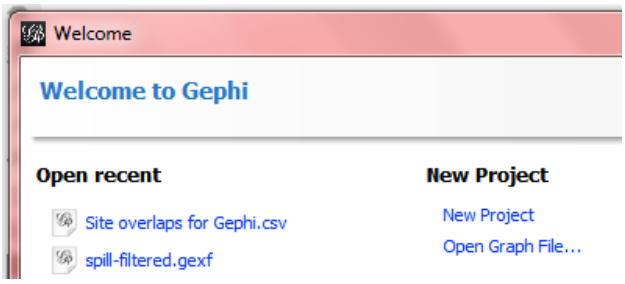
- Open Excel (or a standard text editor)
  - Paste into Excel
  - Remove Null column
  - Remove first header row
  - Edit column headers: Source, Target, mention, retweet, Weight



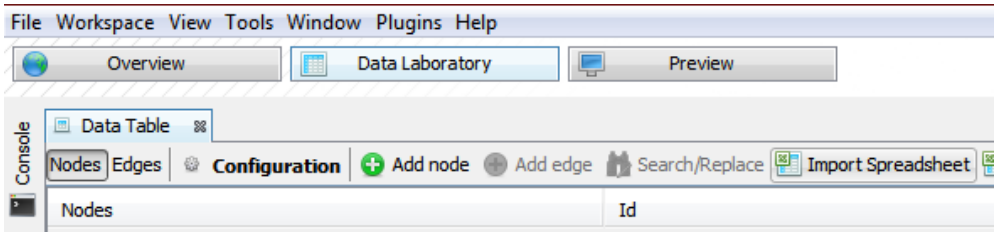
  - Save as CSV

9. Importing data into Gephi

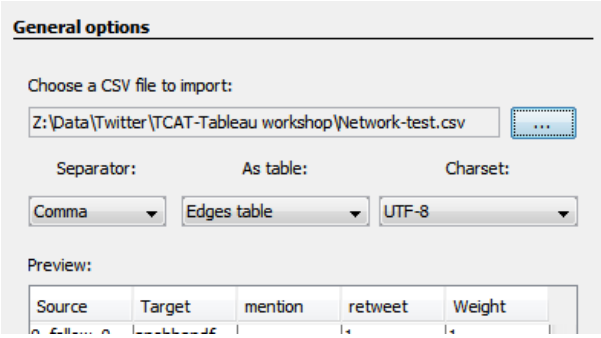
- Open Gephi
- Start a New Project



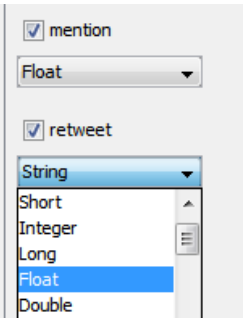
- Go to Data Laboratory tab, Import Spreadsheet



- Select CSV file, choose correct settings (comma, edges table, UTF-8)



- Switch mention and retweet to Float field type, click Finish to import (and wait some time)

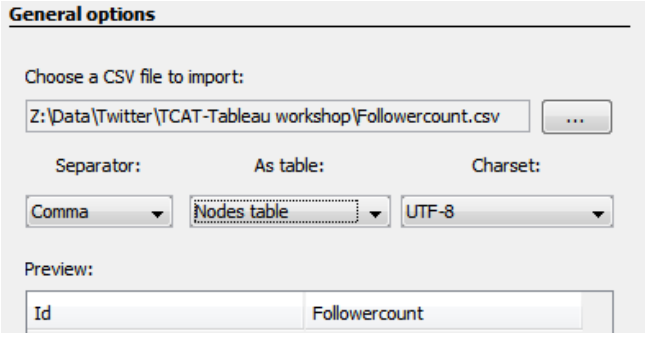


- Process graph as appropriate

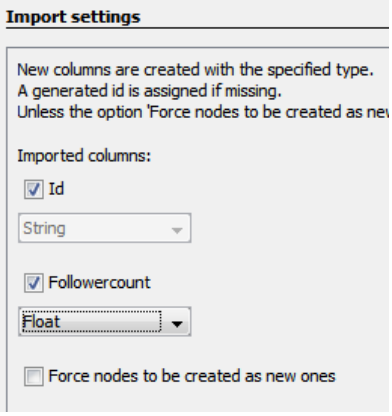
10. Importing additional node data into Gephi

- Exporting from Tableau:
  - Create data tables in Tableau as required – e.g. From User Name, From User Followercount (continue to use LOWER for user names, and use MAX rather than SUM as aggregation for follower count and similar values!)
  - Export via Worksheet > Copy > Crosstab
  - Paste into Excel, rename column headers
    - Note: any username columns should be called Id
  - Export from Excel as CSV

- Importing to Gephi:
  - In the Data Laboratory, use Import Spreadsheet
  - Import the new CSV as a Nodes table:



  - Change any numerical values to Float field type, and don't force nodes to be created as new ones



  - Note that this may have imported additional accounts that were not part of the original network (e.g. users in the original dataset who did not @mention others or were @mentioned by others and were therefore not part of the network dataset). These can be identified and excluded in Gephi using their Degree values.