Step-by-Step Guide to Workshop Activities

This guide documents the analysis steps demonstrated in the Social Media Analytics workshop during the <u>QUT Digital Media Research Centre</u>'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 (<u>a.bruns@qut.edu.au</u>).

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 <u>Twitter Capture and Analysis Toolkit (TCAT</u>), 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 <u>Tableau</u> – 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):

(Paris_2015_UN_Cli	mate_C 🔯 🗙					
	Paris_2015_UN_Climate_Change_Co_Info_re-20150921-20150927fullExport1a7ce924c6.csv ×						
	Does the file include field names in the first row?						
	Yes, the firs	t row has field names in it.					
	No, automa	tically generate names for the fields.					
_	Field Separator:	Comma	ŀ				
Ħ	Text Qualifier:	•					
1	Character Set:	UTF-8					
Ic	Locale:	English (Australia)	ŀ				

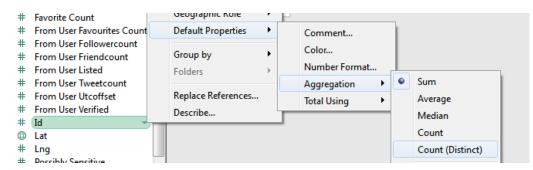
- Drag hashtag and mentions files into whitespace next to main file
- Change relationship to left join, and check file format settings; in each case, join on Id = Tweet Id:

Paris_	2015_UN_Climate_Ch	nange	-00	Paris_2	015_UN_Climate	_Chan	ge
	Join				>	<	
	Inner	Left		Right	Full Outer		
	Data So	ource		Paris_2015_U	N_Climate_Ch		
	Id		=	Tweet Id		_	
■	Add new join daus	æ					
id	Time creat		Hom	osermanie	IEXL		Filter Le
Paris_2	2015_UN_Climate_Ch	ange		Paris_2	015_UN_Climate	e_Char	ige
		L.		Paris_2	015_UN_Climate	_Char	ige
	Join				3	×	
	Inner	Left		Right	Full Outer		
	Data So			2	N_Climate_Ch		
	Id	urce	=	tweet id (Paris			
	Tir Add new join claus	e	_	circeria (rana	2010 011		Filter Le
# Р	#					2	Abc Pa

- 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

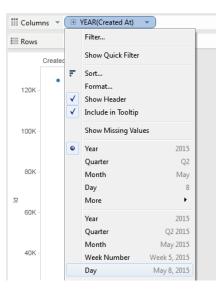
• Move Id field to Measures, change default aggregation to Count (Distinct): CNTD(Id)



- 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

iii Columns	YEAR(Created At)
Rows	CNTD(Id)

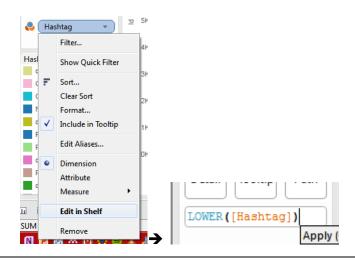
• The default scale for date fields is Year, change this to the second Day option



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

				Sort [Hashtag]	×
				Sort order C Ascending	
UN_Clima Detail Tooltip Path 6K		, i	Descending Sort by		
Hashtag		<u>⊤</u> 5K	Data source order Aphabetic		
ndcount d		Hashtag	4K	Field Id	Aggregation:
etcount offset		Edit Colors	3К	Manual	
ied	 Image: A start of the start of	Format Legends Show Title	2K	Null In IA	Down
ve	 ✓ 	Edit Title Highlight Selected Items	1K	Ide ICLIC IEarth Bearttree	
	F	Sort	0K	Jiken Luke Home Love Jiou	Ŧ
ords	×	Hide Card	→ ¹	Clear OK C	ancel Apply

• 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:



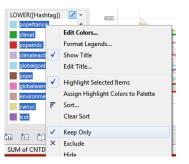
• Exclude generic terms from the list as required: select and right-click



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

	1		👕 View Data: Sheet 1			
			Show aliases	Сору		Export All
			Day of Created At	LOWER([Hashtag])	Id	
			24 September 2015	popeindc	1,736	
	✓ Keep Only 🗙 Exclude 📎 🕶 🏢					
1	LOWER([Hashtag]): popeindc View Data					
/	Day of Created At: 24 September 2015					
	ld: 1,736		Summary Underlyin	g		1 rows
		-	-			

• 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)

	Filter			Filter	
	Show Quick Filter		1	Show Quick Filter	
✓ ✓	Format Show Header Include in Tooltip	/	× ×	Format Show Header Include in Tooltip	
•	Dimension Attribute Measure (Count (Distinct))			Dimension Attribute Measure (Count (Distinct))	
•	Discrete Continuous		•	Discrete Continuous Edit in Shelf	
	Edit in Shelf			Compute using	Table (Across)
Δ	Add Table Calculation		Δ	Edit Table Calculation	Table (Down)
	Quick Table Calculation	Running Total		Clear Table Calculation	Table
	Remove	Difference		Quick Table Calculation	Cell
		Percent Difference Percent of Total		Remove	Created At LOWER([Hashtag])

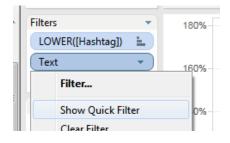
• Switch graph style to to continuous area map



- 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)

4		^	Filters		180%	
	Created At Abc Filter Level Tom User Created At		Add to Sheet Show Quick Filter		160%	
	Abc From User Descriptio # From User Id Abc From User Lang	1	Duplicate Rename			_
	Abc From User Name Abc From User Profile I		Hide		140%	US ET Date
	Abc From User Realname		Create	•	Calculated Field	
	Abc From User Timezone Abc From User Url		Transform	•	Group	DATEADD('hour',-14,[Created At])
	Abc From User Withheld # In Reply To Status Id		Convert to Continuous Change Data Type	•	Set Parameter	DATEADD(date_part, interval, date)
h4			2 71	100		

• Create text filter as quick filter



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

-	-
	Lowercase tweet text
	LOWER([Text])

4. Exploring tweet and retweet texts, using quick filters

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

Filter [User To Name]
General Wildcard Condition Top
Select from list O Custom value list O Use
Enter Text to Search
🔲 Null
🔲 Oboy
ODanSmith
1anhaga
1AshleyMorrison
1aurashep
1 1 1 ClickSolar
1climatejustice
1djcharmain
1ebk
1Heart1Tree
All None

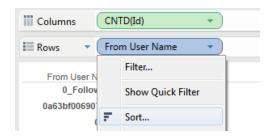
 Text, From User Name, User To Name, Mention Type into Rows – select Add all members if asked

Pages	Columns				
	Rows	Text	From User Name	User To Name	Mention Type
Filters					
User To Name					

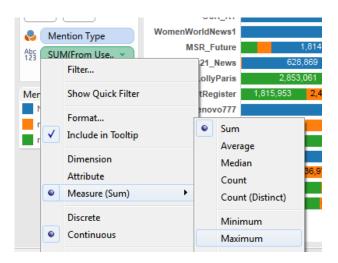
- Turn User To Name filter into quick filter
- Explore:
 - a) e.g. all CNN / BBC / ... accounts
 - b) check (or filter) for retweets
 - c) Explore other variations

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
 - o 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

Dashboard +	Most Active Users			Most Visible Use	PC		Mention Type
ul Sheet 1	MOST ACTIVE OSETS			MOSE VISIBLE OSE	15	1	Null
uli Sheet 2	From User Name			User To Name			mention
Most Active Users	MercianRockyRex	763	~	COP21		-	retweet
Most Visible Users	ClimateWise2015 4	,929 <mark>4,926</mark>		Pontifex			Telweel
	YourGreenNews 627	628		UNFCCC			
	KiraOnClimate 2,926	2,926		CFigueres			

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

Das	hboard Story Analys	sis
ŧ	New Dashboard	e
	Format	Jse
	Copy Image	
	Export Image	Rex
	Cl	1.27

7

6. Geolocation

• Change Lng field to Longitude type

nuc meusure numes	Change Data Type	,	0.1	
Measures	Geographic Role	•	•	None
# From User Friendcount	Default Properties	⊁		Area Code
 From User Listed From User Tweetcount 	Group by	•		CBSA/MSA
# From User Utcoffset	Folders	F		City
# From User Verified				Congressional District
# Id	Replace References			Country/Region
Lat	Describe			County
# (Lng		-		county
# Possibly Sensitive	E			Latitude
# Retweet Count				Longitude
# Time				Chata (Danuin an

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

Columns	Lat
Rows	Lng

• Change to symbol map, explore

	123	123	- 1	10.00	11	-	322	12.92
	123		- 3	1 × 10	10	-	1843	11.00
*	123	123	- 1			1	1142	001

• Note: usually inconclusive due to lack of volume

7. User Percentiles

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

_	I	
Sender Rar	king	
RANK PE	RCENTILE (COU	NTD([Id])

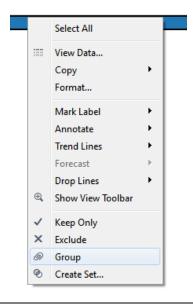
 Graph CNTD(Id) against From User Name, Sender Ranking on Label, order descending

Pages	Columns	CNTD(Id)		
	Rows	From User Name	F	
Filters	From User N	ame		
	MercianRocky	Rex		
	ClimateWise	2015		
G. 1	YourGreenN	ews		
Marks	KiraOnClin	nate		0.99
🖬 Automatic 🛛 👻	joolsmag	ools		0.9999
	tjwise	man		0.9999
Color Size Label	AnimalRight	sJen		0.9999
	liveco	op21		0.9999
Detail Tooltip	RobinBall	1961		0.9999
Detail Tooltip	CSR	RT	0.9	9998
Abc Sender Ranking Δ	WomenWorldNe	ws1	0.99	998
	MSR_Fu	ture	0.999	98
	COP21_N	ews	0.999	8

- Filter by Sender Ranking (repeat twice):
 - 1. Sender Ranking >.99

Range of values	⊫ At least	At most	Special
Range of values			
0.99		1	
0.710256994			1
Reset		OK Can	Include Null Values Apply

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



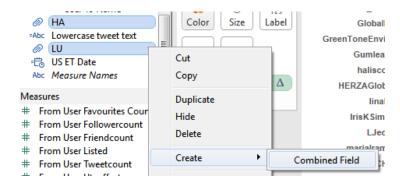
- Edit the new group:
 - o change field name to LU
 - Use Rename button to rename first list to LU (lead users)

Edit Group [l	From User Name (group)]
Field Name:	LU
Groups:	Add to: _Law_of_Fashion, _RichardGoode, _RoyBatty_ anc 🔻
	aw of Fashion. RichardGoode. RovBatty and 590 more.
	JICI
Group	Rename Ungroup 🗹 Show Add Location
V Include	Other' Find >>
Reset	OK Cancel Apply

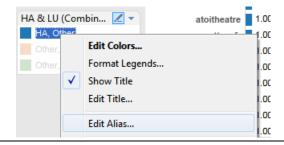
• Remove new LU group from Color

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



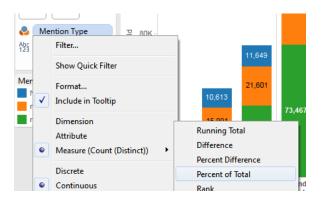
- o Drag combined field onto Color to demonstrate group membership
 - Note: Sender Ranking values are now calculated separately for each category!
- o Use Edit Alias to rename categories in colour legend



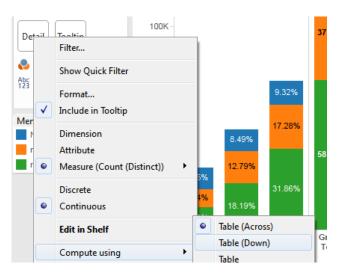
- Using percentile groups in the analysis:
 - o Place combined field on Rows, CNTD(Id) on Columns
 - Use Analysis > Totals > Show Row Grand Totals

Ana	Ilysis Map Format	Server	Window	/ Help
~	Show Mark Labels Aggregate Measures Stack Marks			▼ Normal ▼ → ∠ ▼ ↓ (Combined) ▲
	View Data Reveal Hidden Data		·	
	Percentage of		•	nbined)
	Totals		•	Show Row Grand Totals
	Forecast		•	Show Column Grand Totals

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



o Compute label percentage by using Table (Down)



- Make same changes for CNTD(Id) in Rows: Percentage of Total, Compute Using Table (Down)
- o Note: values above 100% due to some tweets being both retweets and @mentions
- o 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:
 - o User To Name on Rows, Combined field on Columns, CNTD(Id) on Text
 - Sort User To Name by CNTD(Id)
 - o Change CNTD(Id) to Percent of total
 - o 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

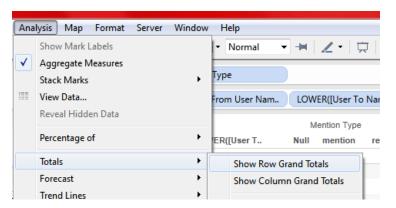
• From User Name, User To Name on Rows, Mention Type on Columns, CNTD(id) on Label

Pages	iii Columr	Mention Type	
	III Rows	From User Name	User To Name
Filters			
Marks Abc Automatic	Drop field here		
Color Size Text			
Abc 123 CNTD(Id)			

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

Columns	Mention Type			
Rows	[LOWER([From User Name])		User To Name	
		Apply (Ctrl+Enter)		

• 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

воок	1		
Wo	rksheet Dashboard	Story Analysis Map Format	Ser
to	New Worksheet	Ctrl+M 🔸 😂 - 🛛 🗗 💺	⁺₹∥
	Сору	Image	hs
	Export	Data Ctrl+C	
	Clear	Crosstab	

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

	A1		(C	∫∞ Sour	ce
2	А	В	С	D	E
L	Source	Target	mention	retweet	Weight
,	0Dan£mit	£10bricho	1		1

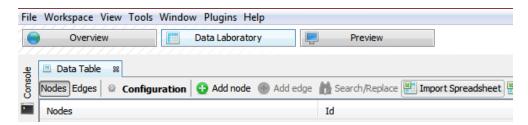
o Save as CSV

9. Importing data into Gephi

- Open Gephi
- Start a New Project



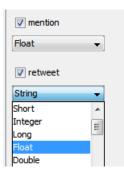
• Go to Data Laboratory tab, Import Spreadsheet



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

General optio	ns				
Choose a CSV	file to import:				
Z:\Data\Twit	ter\TCAT-Tabl	eau workshop	Network-test.	csv)
Separator	:	As table:		Charset:	
Comma	▼ Edges	table	▼ UTF-8	•	
Preview:					
Source	Target	mention	retweet	Weight]
0 4-11 0		1	l	4	

• 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
 - o Paste into Excel, rename column headers
 - Note: any username columns should be called Id
 - o Export from Excel as CSV

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

General options					
Choose a CSV file to) import:				
Z: \Data\Twitter \TC	AT-Tableau workshop\Folk	owercount.csv			
Separator:	As table:	Charset:			
Comma 👻	Nodes table VITF-8				
Preview:					
Id	Followerc	ount			

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

Import settings			
New columns are created with the specified type. A generated id is assigned if missing. Unless the option 'Force nodes to be created as new			
Imported columns:			
Id 🔽			
String			
V Followercount			
Float			
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.