Research Using Twitter Social Media Data

Twitter is an extensive source of social media data, and unusual in that it has both an extensive data sharing infrastructure and an explicitly permissive Terms of Service. This tutorial document shows how to social scientists can use Twitter as a source of research data, without any technical background or programming ability. It shows how to use commonly available data capture and analysis tools, and the kinds of research methods and investigations that they can support.

Note on Software Required for this Tutorial

To follow this tutorial, you will need to use the Chrome browser to access Twitter's web site. The data capture tool (Web Data RA) is a Chrome browser extension that interprets the Twitter Web pages and captures them as spreadsheet data that you can paste into Microsoft Excel (Windows or Mac). You will also use Gephi, an open source network analysis and visualisation application which you can download and install from `gephi.org`. This expects a recent version of the Java Runtime Environment (JRE) to be installed on your machine.

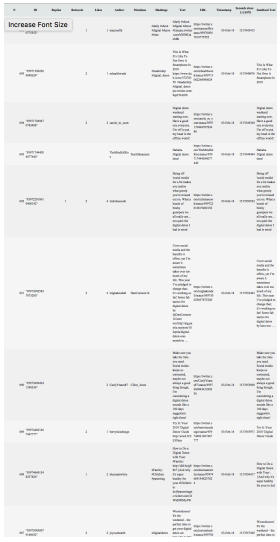
Installing and Using Web Data RA

The Web Data RA will capture Twitter, Facebook and Google data from a browser and allow you to paste a table of information directly into a spreadsheet. This document focuses on its use with Twitter.

1) Install the Web Data RA browser extension into Chrome by visiting `bit.ly/WebDataRA` in Chrome, and clicking on the blue "+ Add to Chrome" button. The small green icon will appear in the top right of the browser window, next to the URL bar.
2) Go to Twitter.com and create a Twitter search or display a timeline
3) Click on the WebDataRA icon to start collecting tweets. Every five seconds the browser will automatically scroll to the bottom of the page to make Twitter load the next batch of results and copy the data automatically to the clipboard.
4) When you have collected enough results, paste the data into an Excel spreadsheet.
5) Use Excel to analyse data, or export to other programs such as Gephi or Voyant for other kinds of analysis.

Overview

When you paste data from WebData RA you will create four tables in your spreadsheet, appearing below each other.



The tweet data, with author, mentions, hashtags, text and counts of retweets, replies and likes broken out in separate columns.

Account	Author Count	...Inc RTs	Mention Count	...Inc RTs
1234margs	1	3		
2202Gayle	1	1		
25Score	1	2		
50Folds	1	3		
90_mile_beach	1	1		
949powerfm	1	2		
ABRAMSbooks	1	5		
ADJBlog	1	2		
AFARmedia			2	6
ASMRLittleMelon	1	2		
ASTsupportAAli	1	1		
AbhaReach	1	2		
AdamtheBallFAN			1	1
AdeldMeyer			1	1
AdrianChen			1	2
AgenceW	1	2		
AlanSee	1	3		
AlexRWilliams	1	4		
AlishaValerie	1	2		
Aloe_Vera_FLP	1	2		
AmlyBotanicals			1	6
Andrea_MacE	1	2		
AngelaInBold			1	1
Anitapratapsing	1	2		
AnnaBell_writes			1	5
Anonfighter5			1	1
AppliedIG	1	4		
Argentine58			2	2
AstraLifeID	1	5		
AvonworthEA	1	2		
BBC			1	1
BBCJonSopel	1	1		
BBCWomansHour			1	2

Account occurrence summary, a count of the number of times that each Twitter account appears in the dataset as author or a mention (including the number of retweets).

Hashtags	Count	...Inc RTs
##skincare	2	2
#124;	1	1
#365daysofselfcare.	1	1
#AI.	1	1
#AageSeRight	1	4
#AbuDhabi	1	2
#Apps	1	2
#Argyll	1	15
#BCNMI	1	1
#BLOGOSPHERECHAT	1	2
#BabyZakTime	1	2
#Bachelor	1	3
#BeWellDoWell	1	5
#BienvenueCheznous	1	2
#BlogosphereChat	15	27
#BlogosphereChat.	1	1
#CBCconf18	1	3
#Cafepress	2	3
#CaribbeanSea!	1	7
#CellPhone	1	7
#Cheshire	2	19
#Children	15	51

Counts of the appearances of each hashtag.

Source	Target	Weight
1234margs	YouTube	1
2202Gayle	BlogosphereM	1
2202Gayle	ccfest	1
2202Gayle	edfringe	1
90_mile_beach	sundarakarma	1
ABRAMSbooks	tanyagoodin	1
ASMRLittleMelon	Anonfighter5	1
Andrea_MacE	iamwellandgood	1
AppliedIG	DanConnors16	1
BBCJonSopel	BBC	1
BBCJonSopel	OilPaul	1
BBCJonSopel	SebGorka	1
BVG_Kampagne	adidas	1
BVG_Kampagne	betablogr	1
BarryEdwardsJr	LloydEdwards91	1
BellyDanceRav		

A table of edges of the conversational network, i.e. the number of times each pair of accounts communicate with each other.

You can use this data in various ways:

(a) The tweet data (gray) contains the basic data about each tweet: what was said, when, by who and to whom. You can use this data to form a general overview of the communication over time and identify the most significant tweets. You can also examine specific tweets and their context by referring back to the Twitter site using each tweet's URL.

(b) The account table (green) shows you the most active tweeters, the most frequent repliers, and the most retweeted users. This will help you see the key actors in a conversation, and the main roles that they take. You can follow up by clicking on the account names to look at the account bios and the relevant timelines of these actors to understand whether they are corporate accounts, private individuals, bots or trolls.

(c) The hashtag table (blue) shows you the most frequently used hashtags. This can help you extend your data gathering to look for more tweets relevant to your research question.

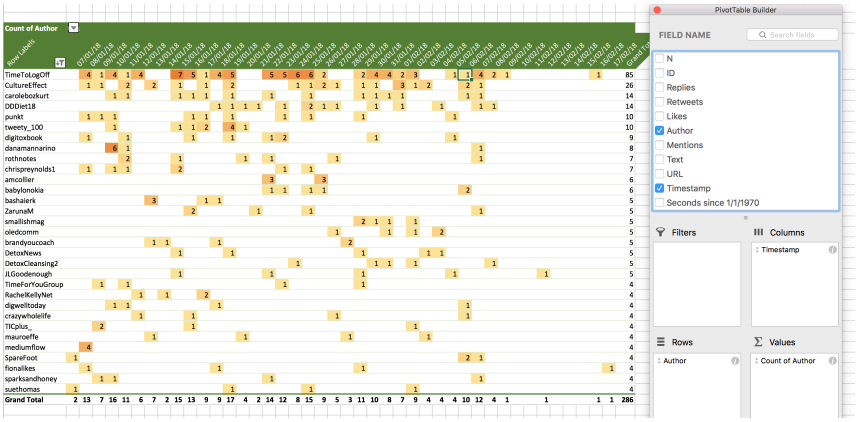
(d) The edge table will help you to see the interactions between actors, and help you to understand groupings of actors, and the pattern of their interaction. Is a key account dominating a conversation and talking to many others? Are they responding or just being passive recipients of marketing messages? Is there a group of equals having a balanced conversation with equal participation?

Using Excel for Simple Quantitative Analyses

The following set of tweets (gray table) comes from a Twitter search for the phrase “Digital Detox”.

N	ID	Replies	Retweets	Likes	Author	Mentions	Hashtags	Text	URL	Timestamp	Seconds	Sanitised Text
58	"959715062		2	1	mhealthweek		#leadership #digital_detox	This Is What It’s Like To Not Own A Smartphone In 2018 https://www.dub.io/tw/37270350 #leadership #digital_detox pic.twitter.com/8cpI70Al0S	https	03-Feb-18	1517648876	This Is What It’s Like To Not Own A Smartphone In 2018
57	"959717045		2	2	carole_m_scott			Digital detox weekend starting now. Have a good one everyone. I’m off to put my head in the offline world!	https	03-Feb-18	1517649349	Digital detox weekend starting now. Have a good one everyone. I’m off to put my head in the offline world!
56	"959722010	1	2	4	nicholascook			Being off 'social media' for a bit makes you realise what gossip you've missed out on. What a bunch of bitchy gossipers we all really are... not quite the digital detox I had in mind	https	03-Feb-18	1517650532	Being off 'social media' for a bit makes you realise what gossip you've missed out on. What a bunch of bitchy gossipers we all really are... not quite the digital detox I had in mind
55	"959730025		2	4	leighakendall	DanConnors16		I love social media and the benefits it offers, yet I’m aware it sometimes takes over too much of my life. This year I’ve pledged to change that; it’s working so far! Some fab tactics for digital detox by @DanConnors16 here too:http://aig.passle.net/post/102ep4a/digital-detox-one-month-in …	https	03-Feb-18	1517652443	I love social media and the benefits it offers, yet I’m aware it sometimes takes over too much of my life. This year I’ve pledged to change that; it’s working so far! Some fab tactics for digital detox by here too: …
54	"959736094			2	CarlyYeates87	CIJoe_Jones		Make sure you take the time you need. Social media keeps us connected, maybe not always a good thing though, I’m considering a digital detox. sounds like a 100 days suggestion right there!	https	03-Feb-18	1517653890	Make sure you take the time you need. Social media keeps us connected, maybe not always a good thing though, I’m considering a digital detox. sounds like a 100 days suggestion right there!
53	"959744821		2	1	berrynicedesign			Try It: Your 2018 'Digital Detox' Guide http://crwd.fr/2E35bey	https	03-Feb-18	1517655971	Try It: Your 2018 'Digital Detox' Guide
52	"959746691		1	2	sheenamwhite		#Family: #Children #parenting	How to Do a Digital Detox with Your #Family: http://dld.bz/g8Rr7 [And why it's super healthy for your #Children to do]#parentingpic.twitter.com/DWId9MMy4W	https	03-Feb-18	1517656417	How to Do a Digital Detox with Your : [And why it's super healthy for your to do]
51	"95975083807914803		2	2	joyoushealth		#digitaldetox	Woooohoooo! It's the weekend – the perfect time to get your digital detox on: http://bit.ly/2Eo1JZR #digitaldetoxpic.twitter.com/bjCKvXqU8F	https	03-Feb-18	1517657405	Woooohoooo! It's the weekend – the perfect time to get your digital detox on:
50	"959754175	3	2	18	griffski			*taps wineglass* folks, i have gathered you here today to something something digital detox gwenyth paltrow going retrograde embracing obsolescence etcetera. i'll be online & on email if you need me,	https	03-Feb-18	1517658201	*taps wineglass* folks, i have gathered you here today to something something digital detox gwenyth paltrow going retrograde embracing obsolescence etcetera. i'll be online

The easiest way to see an overview of a Twitter timeline is to create a Pivot Table. Click on any gray cell, and choose “Pivot Table” from the Insert ribbon. In the Pivot Table builder, drag “Author” from the Field Name panel into the “Rows” panel, drag “Timestamp” into the “Columns” panel, and drag “Author” (again) into the “Values” panel (it will automatically turn into “Count of Author”).



The screenshot above shows the dates included in the twitter sample as green column headings, and the accounts that authored tweets as row labels along the left hand side, ordered by most prolific tweeters. The values in each cell are the number of tweets authored by a specific account on a specific day.

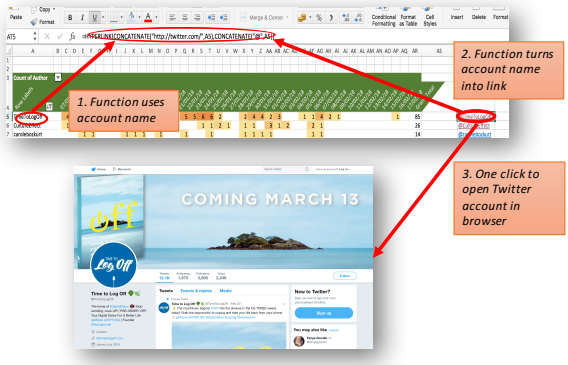
You can adjust the formatting for convenience (I narrowed the columns and slanted the column headings and changed the angle of the text to 60 degrees to fit), use the “Row

Labels" control to sort by the author count (*i.e.* the number of tweets an author created) and show only the rows where the total author count is greater than a chosen threshold. You can also use conditional formatting to colour the cells to highlight the most extreme values.

All kinds of summaries and analyses are possible using Excel on this data, including:

- Showing the distribution of the tweet sample through time
- Identifying the most prolific and/or popular actors, and showing their activity through time
- Showing the use of individual hashtags (this might be useful in a big conversation, or one that evolves over a longer period)
- Comparing the relative proportion of contributions from different actors / hashtags

All of these analyses will lead on to other questions that can be asked by going back to Twitter.



The account names in the account "author and mentions" (green) table are clickable, and open the page of the account profile in your default web brower.

Alternatively, to make a set of account names into clickable hyperlinks giving browser access to the user's Twitter timeline and bio, use the following function:
`=HYPERLINK(CONCATENATE("http://twitter.com/",A9),CONCATENATE("@",A9))`
where A9 is an example of a cell address that contains a Twitter account name e.g. lescarr.

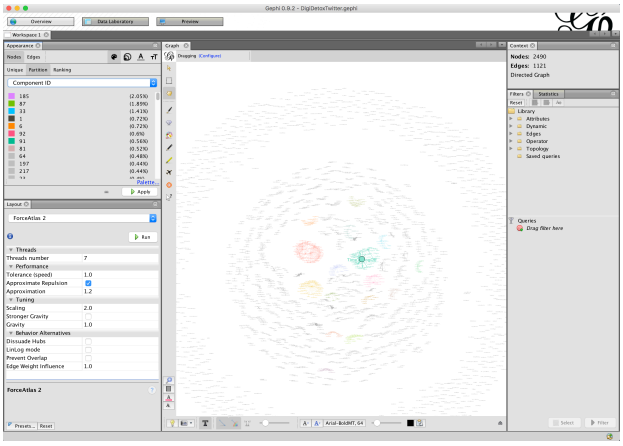
Following the account hyperlinks for the most prolific authors in the green table, we see that they are all commercial actors to one extent or another.

Account	#	Bio
ItsTimeToLogOff	30	Time To Log Off is the home of digital detox. We're spearheading the movement to disconnect regularly from digital devices and reconnect with the world offline. We do this through collecting facts on the need for digital detox, running campaigns to get everyone off their screens and hosting retreats, events and workshops.
DinnerTableMBA	9	A commercial organisation working together to help families become more confident, successful, and self-empowered
SpareFoot	8	A storage company. We make it easy to move and store your stuff. Reserve storage for free and get your mind out of the clutter.
CultureEffect	5	Author of Digitox: How to Find a Healthy Balance for your Family's Digital Diet

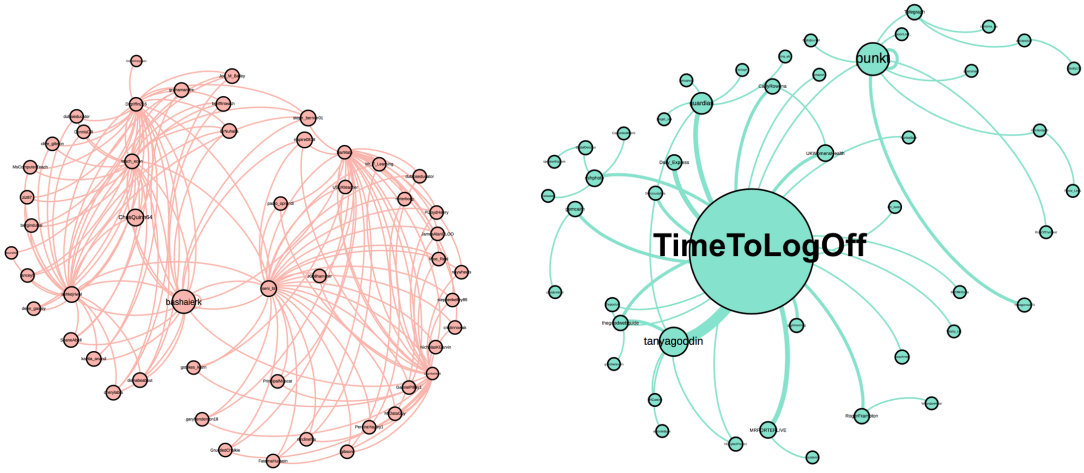
Using Gephi for Network Analyses

To see how the various accounts interact with each other as a network, copy and paste the yellow table into a separate spreadsheet and save it as a CSV file (call it `edgetable.csv` or

similar). Load up the network visualisation program “Gephi”, and start a new project. In the “Data Laboratory”, choose “Import Spreadsheet” and load up the CSV data as an edge table. You can then apply a variety of network layout algorithms in the “Overview” pane.



The network visualisation shows many isolated nodes (accounts) in an outer ring and a central core made up of different groups of accounts. Many of these are loosely connected “chains” of 2-6 accounts where one account has mentioned another, which has mentioned another and so on. There are more complicated subnetwork components that demonstrate more activity, as seen below.



The green component is dominated by a single corporate account (the most prolific account in this sample) whose role is to promote the idea of a digital detox and that “tweets at” many other accounts, initiating communication with them. By contrast, the red network consists of a larger group of teachers and education professionals who already participate in a larger professional network within Twitter, and who are discussing the topic of digital detox within that context.

Many summaries and analyses are possible using Gephi’s network visualisation tool:

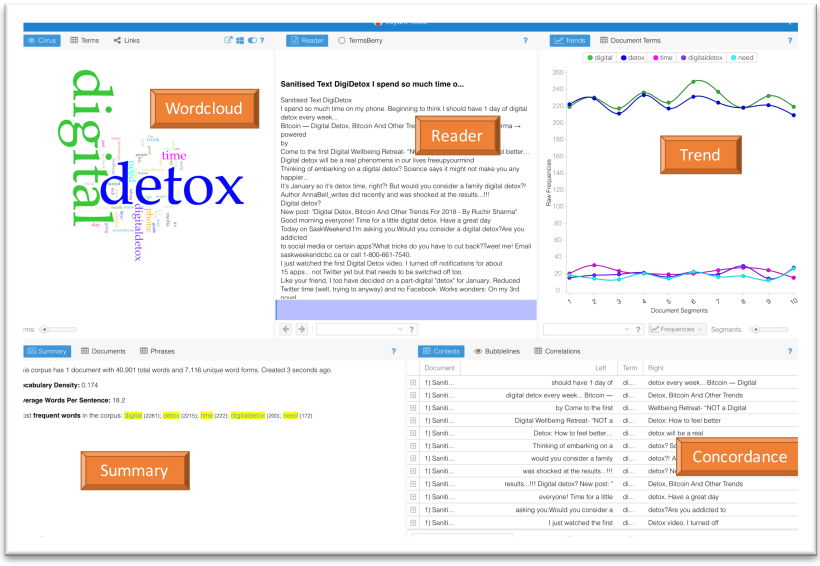
- Showing the interaction of the network actors
- Identifying the communities and active participant subgroups within the larger sample
- Identifying the roles of different actors in the communications network

***Please note**, that many network analyses in the literature often focus on the retweet network, but that is not possible with Web Data RA, because the identity of the retweeting accounts is not available to the Web browser (the Web User interface or Web app). You have data on the number of retweets (i.e. the popularity of the tweet), but not the accounts that retweeted the original message. For more information on a supplementary service being developed to fill this gap, see the end of this document.*

Using Voyant for Textual Analyses

In the gray table, copy the "Sanitised Text" column. This contains the text of all the texts, but with all the Twitter features (@names, #hashtags, URLs) removed to leave only the English text.

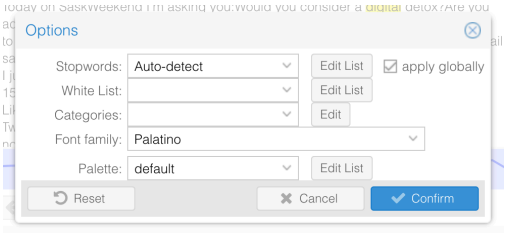
Go to the Voyant-Tools.org website, paste the text into the textbox and press the "Reveal" button. You will see a screen with several panels that help you explore the text of the tweets in different ways. Voyant Tools is a textual corpus analyser. It considers the data that you have entered as a single document where the individual tweets are like individual sentences or paragraph.

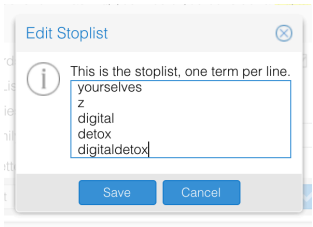


The Wordcloud, Reader, Trend and Concordance all analyse the text from the collection of tweets. Click on a word in the Wordcloud, and all its occurrences will be highlighted in the Reader panel, it's frequency throughout the whole document (set of tweets) will be displayed in the Trend graph, and its context will be displayed in the Concordance. This helps you to investigate the use of language in the collection of tweets, and quickly understand what is being talked about and how. It also helps you to see how the language changes over time. The first thing that this display shows is that the most common terms are *digital*, *detox*, and *digitaldetox* because they were the search terms! To ignore them, add them to the stop words list by clicking on the "Options" icon in the Wordcloud panel's grey icon bar.

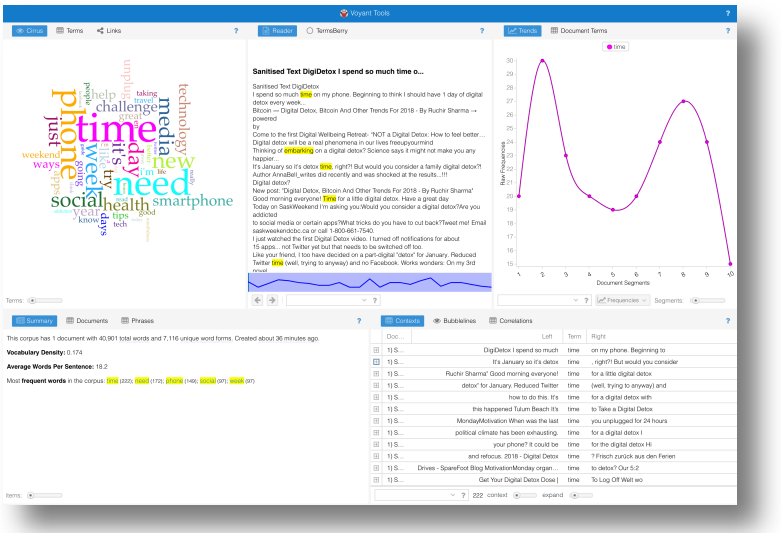
When the mouse enters the bar, you will see the following icons appear:  .
The "Options" icon is the sliding button, next to the "Help" question mark.

To edit the Stopwords list, click on the "Edit List" button and add the three extra terms to be ignored, one per line.



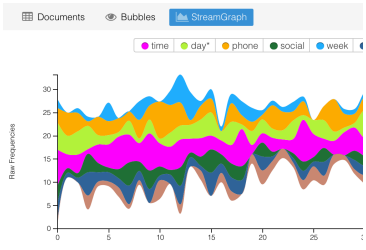


You will see that the panels have re-analysed the text, ignoring the terms that you have added to the stopword list.

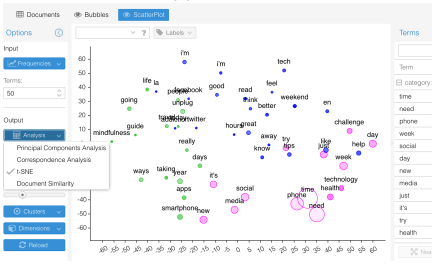


Other more complex visualisations and analyses are available to be used, including advanced Machine Learning algorithms to cluster keywords and simple graph visualisations to show keyword co-occurrence.

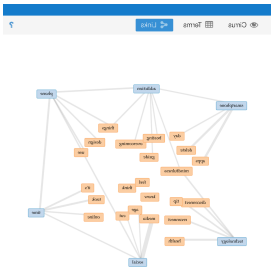
Stream Graph



Dimensional Reductions of Keyword Appearance



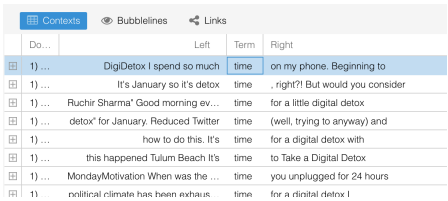
Network of Keyword Co-occurrences



Analysis of Twitter Language

The word cloud shows in visual form the most commonly used terms in the Twitter sample. This is great for an impression of the topics, but a more useful summary is the "Terms" panel.

Using 15 of the top 20 words in this sample, we might hypothesise that a digital detox is about spending less **time** using a **phone** and the **need** to **unplug** yourself from **smartphone social media technology apps** – it's a **health challenge** you **try** for a **day**, a **week** or a **year**.



An easy way to investigate the use of each of these words is to select them in the term panel, and scroll through the in-context use in the "Context" panel. The most commonly used word is "time" but it's main use is not in the sense of "spending too much time" or "saving time" but in the sense of an opportunity (it's time to… or it's time for…) and frequently as a rhetorical question (Isn't it time for a digital detox?)

Finding tweets that speak from personal experience

Many of the tweets seem to be promotional lifestyle tweets in headline form (*e.g.* "Why You Should Do a Digital Detox" or "Digital Detox Benefits: Are you addicted to technology?") To identify personal tweets from people who have tried or are thinking of trying a digital detox, search for tweets containing the personal pronoun "I".

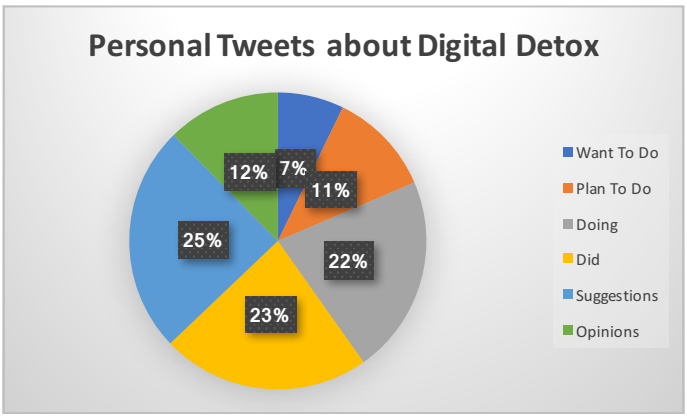
Unfortunately, "I" is one of the Voyant Tools stoplist words and is ignored (a stoplist consists of the common but low-information words in a text including pronouns, prepositions and conjunctions). Although it is possible to edit the stoplist in Voyant, I will search for the string *I<space>* or *I<apostrophe>* in the text of the tweet (to match I, I am, I have, I will, I'm, I've, I'll).

Use the following formula

```
=OR(NOT(ISERROR(FIND("I ",A1))),NOT(ISERROR(FIND("I'",A1))))
```

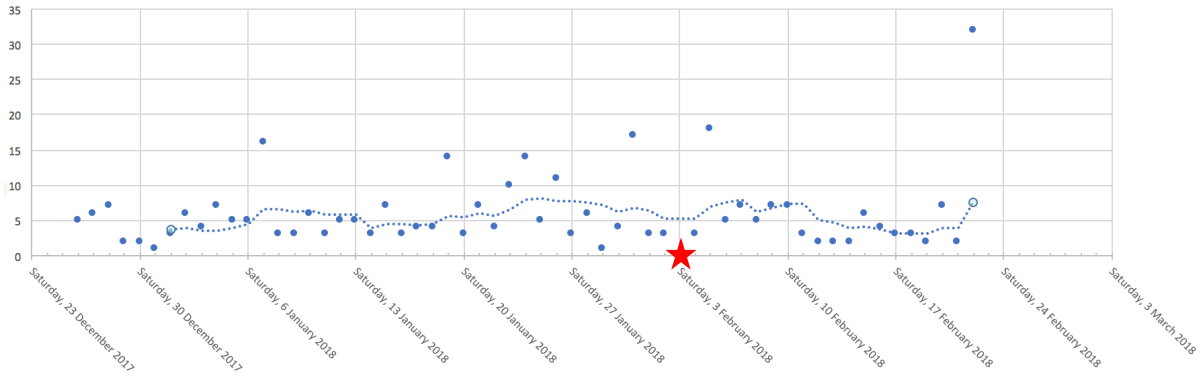
to define an extra column in the gray table, and sort the set by the results (TRUE or FALSE). Only 17% of the tweets in this collection were identified as personal by this simple rule. (Note, not every tweet with the word I is talking from a first person perspective, and many tweeters would miss "I have" or "I am" from the beginning of a tweet or text.)

Of those 17%, just over half were actively planning a digital detox, were doing a digital detox or had done one. (The irony of tweeting about doing a digital detox is not lost on some commentators!)



It is then possible to follow up with each of the accounts that has tweeted about an actual period of digital detox that they have undertaken, to examine their Twitter activity before, during and after the detox and to identify any quantitative difference in their use of the platform.

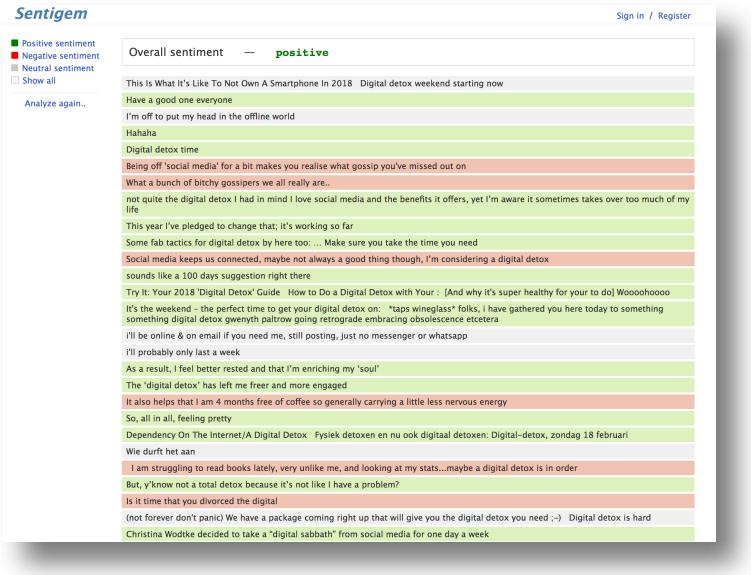
As an example, I chose a single account that had announced that they were starting a "Digital Detox" to deprive themselves of social media for a day. I accessed all the tweets of that user over the period (using WebDataRA), and created a 1-author pivot table (as above) to aggregate the tweets to a count per day, and then plotted that data as a scatter graph with 7-day moving average. The graph shows that although the account did not tweet on the chosen date (marked with a red cross) there is no noticeable change in tweeting behaviour after the announced detox day.



Additional Methods: Sentiment Analysis

Sentiment analysis can help you identify positive or negative comments in your sample. This is a popular method in industry, especially with brand management companies. However it is academically contested, and does not have a high degree of transparency in the lexical processing.

Nevertheless, to try it out, paste the "Santitised Text" column into an online service such as `sentigem.com`. Consider to what extent the results seem accurate to you – how well does it identify positive and negative 'sentiment' in a sentence? What kinds of inaccuracies can you see? And, most importantly, despite any shortcomings, does it help you to identify any points of interest in your data for more thorough investigation?



Additional Methods: Retweet Network

A supporting service is being developed for WebDataRA to allow you to recover the retweet network around the tweets that you have collected. As previously explained, the Twitter web app does not show the retweets of each tweet, only the count of retweets. However, it is possible to request that data from the Twitter API, although it is limited to 100 retweets of each tweet. (If a tweet has been retweeted more than 100 times, those extra tweets will not be accessible, as such this should be considered to be an approximation of the real retweet network.)

To obtain the retweet data for your tweets, select the twitter ids for which you wish to find the retweets, and paste these into the form at http://pretend.webdataRA /retweets/ . *Please note that the Twitter API limits requests such that each twitter ID that you provide will take 12 seconds to process, and the finished result may take an hour or more.* You can paste the resulting data into an Excel spreadsheet, save it as CSV and import it into Gephi as previously explained.

In the Gephi network visualisation below, nodes are larger and coloured more intensely according to how much other nodes retweet them (their role as authorities in the network), but the node labels are sized according to whether they retweet or are retweeted. So you can see that some accounts are very active *but not as originators of information*, whereas other accounts may be less active *but are more influential*.

RT Source	RT Target	Weight
111publishing	Siegel_Jan	1
1georgerichmond	samantha_swift1	1
1queenbbw	IntgrtFamSrvc	1
21flavofsplendr	Rapt_Motherhood	1
23codestreet	anisahob	1
4everK_Hamm	WellnessGSU	1
ADSuthar21	eshagupta2811	1
ALIVEnuigalway	ForoigeMayo	1
Aaawara_Hun	hvgoenka	1
AajizMazhar	JensRoehrich	1
Adamhowells99	ITVBe	1
AdarshAdsati28	hvgoenka	1
AliSasongko	mashable	1
Amaliah_Tweets	anisahob	1
AmanJalan7	hvgoenka	1
AmandaLWaldrop	BlogosphereM	1
AmdyFall	chandraosborn	1
AndrewHill10	mihphoto	1
AnkitPa58532812	OberoiHotels	1
AnqiKrug	aifonline	1
AravindMaveric	eshagupta2811	1
ArgyllSeaGlass	econaturehols	1
Argyll_IslesApp	econaturehols	1
Ashok88189858	eshagupta2811	1
AventurandoSt	lionwhispererSA	1
Aygg92	keemanxp	1
AygunMe94919428	Eylul__Dilan	1
BALLANDCO	spotlightstat	2
BMClaypool	iamwellandgood	1
BMTMRAC	ICTKensington	1
BanjacJasna	VisitCheshire	1
BarreroStadl	jjcanocortes	1
BenAcheson	FastCompany	1