Critically engaging with social media research tools

Dr Steven McDermott
s.mcdermott@lcc.arts.ac.uk
@soci
https://snacda.com
• Dr Steven McDermott, London College of Communications, University of the Arts London.

• TAGS; Yourtwapperkeeper; DMI-TCAT; Gephi; Leximancer; for collecting, archiving, analysing, visualising and disseminating social data collected from Twitter, YouTube and text based data.
Surveillance Capitalism

“The ugly truth here is that much of “big data” is plucked from our lives without our knowledge or informed consent. It is the fruit of a rich array of surveillance practices designed to be invisible and undetectable as we make our way across the virtual and real worlds.” (Zuboff, 2016)
The goal of big data analytics is to change people’s behaviour at scale.

A Chief Data Scientist of a Silicon Valley company that develops applications to improve students’ learning states that...

“The goal of everything we do is to change people’s actual behavior at scale. When people use our app, we can capture their behaviors, identify good and bad behaviors, and develop ways to reward the good and punish the bad. We can test how actionable our cues are for them and how profitable for us.” (Zuboff, 2016)
Future Research Plans
Choose Your Targets Carefully

• Big Data Scientists as Knowledge Creators - *An Ethnography of Data Scientists of 2016*. A searchable and interactive network graph of Data Scientists archived in February 2016 - the digital environment and how it affects individuals and societal groups behaviour in the way they take decisions and seek information.


• Members of Parliament in the Digital Environment - How are members of the political class exchanging information *UK MPs on Twitter August 2014*
Get TAGS

To start using TAGS select one of the versions below then follow the steps below.

**TAGS v6.0**

**TAGS v6.1**

Which version? If you've not used TAGS before I recommend TAGS v6.1 which has an easy setup. If you've setup TAGS v6.0 you can keep on using that version, your existing archives will keep collecting tweets using your existing authentication. Some background in this post.

1. After your copy has been made open TAGS > Setup Twitter Access and follow the onscreen instructions (when selecting this option you'll be promoted to authorize the script to run several services).

   **Important:** In the new version of TAGS you only need to run the setup once

2. Enter what data you want to collect on the Readme/Settings sheet and hit TAGS > Run Now!

What's new

By Martin Hawksey
Twitter Authorisation

Please be aware that if you are logged into multiple Google accounts your access details will be saved with you default account.

By clicking 'Sign in with Twitter' you agree to abide by the Twitter Policy on 3rd party access
<table>
<thead>
<tr>
<th>Archive ID</th>
<th>Keyword / Hashtag</th>
<th>Description</th>
<th>Tags</th>
<th>Screen Name</th>
<th>Count</th>
<th>Create Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>#bigdata</td>
<td>preliminary scrape</td>
<td>research, soi</td>
<td></td>
<td>1040644</td>
<td>Tue, 30 Jun 2015 12:39:44 +0000</td>
</tr>
<tr>
<td>23</td>
<td>#analytics</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>403748</td>
<td>Tue, 30 Jun 2015 13:17:30 +0000</td>
</tr>
<tr>
<td>24</td>
<td>#lot</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>1051596</td>
<td>Tue, 30 Jun 2015 13:17:47 +0000</td>
</tr>
<tr>
<td>25</td>
<td>#datascience</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>179621</td>
<td>Tue, 30 Jun 2015 13:18:14 +0000</td>
</tr>
<tr>
<td>26</td>
<td>#deeplearning</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>113215</td>
<td>Tue, 30 Jun 2015 13:18:47 +0000</td>
</tr>
<tr>
<td>27</td>
<td>#data</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>463725</td>
<td>Tue, 30 Jun 2015 13:19:06 +0000</td>
</tr>
<tr>
<td>28</td>
<td>#machinelearning</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>139813</td>
<td>Tue, 30 Jun 2015 13:20:25 +0000</td>
</tr>
<tr>
<td>29</td>
<td>#internetofthings</td>
<td>secondary scrape</td>
<td>research, soi</td>
<td></td>
<td>169557</td>
<td>Tue, 30 Jun 2015 17:57:39 +0000</td>
</tr>
</tbody>
</table>
## Bills

**Date:** August 2016

### Summary

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS Service Charges</td>
<td>$55.53</td>
</tr>
<tr>
<td>Usage Charges and Recurring Fees</td>
<td>$55.53</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$55.53</td>
</tr>
</tbody>
</table>

### Other Details

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payment Summary</td>
<td>$55.53</td>
</tr>
<tr>
<td>Tax Invoices</td>
<td>$55.53</td>
</tr>
</tbody>
</table>

### Details

<table>
<thead>
<tr>
<th>Description</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS Service Charges</td>
<td>$55.53</td>
</tr>
<tr>
<td>Data Transfer</td>
<td>$0.05</td>
</tr>
<tr>
<td>Elastic Compute Cloud</td>
<td>$46.24</td>
</tr>
<tr>
<td>CT to be collected</td>
<td>$0.00</td>
</tr>
<tr>
<td>GST to be collected</td>
<td>$0.00</td>
</tr>
<tr>
<td>US Sales Tax to be collected</td>
<td>$0.00</td>
</tr>
<tr>
<td>VAT to be collected</td>
<td>$9.24</td>
</tr>
</tbody>
</table>
### Amazon Web Services

#### Compute
- EC2: Virtual Servers in the Cloud
- EC2 Container Service
- Run and Manage Docker Containers
- Elastic Beanstalk
- Run and Manage Web Apps
- Lambda: Run Code without Thinking about Servers

#### Storage & Content Delivery
- S3: Scalable Storage in the Cloud
- CloudFront: Global Content Delivery Network
- Elastic File System
- Glacier: Archive Storage in the Cloud
- Snowball: Large Scale Data Transport
- Storage Gateway: Hybrid Storage Integration

#### Database
- RDS: Managed Relational Database Service
- DynamoDB: Managed NoSQL Database
- ElastiCache: In-Memory Cache
- Redshift: Fast, Simple, Cost-Effective Data Warehousing
- DMS: Managed Database Migration Service

#### Developer Tools
- CodeCommit: Store Code in Private Git Repositories
- CodeDeploy: Automate Code Deployments
- CodePipeline: Release Software using Continuous Delivery

#### Management Tools
- CloudWatch: Monitor Resources and Applications
- CloudFormation: Create and Manage Resources with Templates
- CloudTrail: Track User Activity and API Usage
- Config: Track Resource Inventory and Changes
- OpsWorks: Automate Operations with Chef
- Service Catalog: Create and Use Standardized Products
- Trusted Advisor: Optimize Performance and Security

#### Internet of Things
- AWS IoT: Connect Devices to the Cloud

#### Game Development
- GameLift: Deploy and Scale Session-based Multiplayer Games

#### Mobile Services
- Mobile Hub: Build, Test, and Monitor Mobile Apps
- Cognito: User Identity and App Data Synchronization
- Device Farm: Test Android, iOS, and Web Apps on Real Devices in the Cloud
- Mobile Analytics: Collect, View, and Export App Analytics
- SNS: Push Notification Service

#### Application Services
- API Gateway: Build, Deploy, and Manage APIs
- AppStream: Low-Latency Application Streaming
- CloudSearch: Managed Search Service
- Elastic Transcoder: Easy-to-Use Scalable Media Transcoding
- SES: Email Sending and Receiving Service
- SNS: Message Queue Service
- SWF: Workflow Service for Coordinating Application Components

#### Security & Identity
- Identity & Access Management: Manage User Access and Encryption Keys
- Directory Service: Host and Manage Active Directory
- Inspector: Analyze Application Security
- WAF: Filter Malicious Web Traffic
- Certificate Manager: Provision, Manage, and Deploy SSL/TLS Certificates

#### Resource Groups
A resource group is a collection of resources that share one or more tags. Create a group for each project, application, or environment in your account.

#### Additional Resources
- Getting Started
- AWS Console Mobile App
- AWS Marketplace
- AWS re:Invent Announcements

#### Service Health
- All services operating normally.
- Updated: Sep 20 2016 13:19:01 GMT+0100
DMI Twitter Capturing and Analysis Toolset (DMI-TCAT)  
By Bernhard Rieder

Data selection

Select the dataset:

digitalhealth --- 149 330 tweets from 2016-06-21 14:30:00 to 2016-08-08 17:32:43

Select parameters:

Query: (empty; containing any text*)
Exclude: (empty; exclude nothing*)
From user: (empty; from any user*)
From twitter client: (empty; from any client*)
(Part of) URL: (empty; any or all URLs*)
Startdate (UTC): 2016-06-21
Enddate (UTC): 2016-08-08

Overview of your selection

Dataset: digitalhealth #digitalhealthcapital #digitalhealth #patient #patients #patientvoice patientvoices digitalhealth mHealth

Search query:
Comments:
Exclude:
From user:
From twitter client:
(Part of) URL:
Startdate: 2016-06-21
Enddate: 2016-08-08
Number of tweets: 149 330
Number of distinct users: 33 050
DMI Twitter Capturing and Analysis Toolset (DMI-TCAT)

Has 36 different options – for extracting or downloading the data that has been collected.

DMI-TCAT is available on Github and is free. However installing it on a server requires money and so does running it.
The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform

Download FREE
Gephi 0.9.1

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos

Support us! We are non-profit. Help us to innovate and empower the community by donating only 8C:

Donate

APPLICATIONS

- Exploratory Data Analysis: intuition-oriented

Like Photoshop™ for graphs.

PAPERS
The Various Algorithms to Choose From…
Social Network Analysis – By John Scott

Social Network Analysis – Methods and Applications

TROPEC
High Performance Text Analysis for Professional Users

Analysis of written or spoken texts requires that certain questions should be asked with regard to certain objectives. To obtain answers to these questions, texts must be reduced as far as possible to their essentials.

Designed for Information Science, Market Research, Sociological Analysis and Scientific studies, Tropes is a Natural Language Processing and Semantic Classification software that guarantees portinence and quality in Text Analysis.

Extraction of Relevant Information

Tropes can immediately detect contexts, isolate themes and identify principal actors, through the application of three levels of semantic classifications. You can quickly determine who says what to whom; who does what; where and when; and with what purpose.

Qualitative Analysis and Categorization

Tropes identifies the Text Style in order to place it in context and rapidly compare it with other texts. Tropes uses Semantic Meta-Categories to group verbs, adjectives, adverbs, personal pronouns and conjunctions.

Chronological Analysis

Tropes carries out a chronological analysis of a text from which the principal episodes can be isolated, the discussion blocks visualized and the development of an idea followed up.
Leximancer School Licensing

Academic Desktop Licence covering all users in a specific School or Faculty. Site-wide licensing options are also available for academic establishments. For more information please contact us.

- Academic School Licence - Please contact us for pricing.

QUALIFIED ACADEMICS AND STUDENTS MAY PURCHASE A SPECIAL, REDUCED COST LEXIMANCER LICENCE FOR INDIVIDUAL, NON-COMMERCIAL USE. THIS OFFER IS NOT VALID FOR COMMERCIAL CUSTOMERS.

- Desktop Licence: Buy
- 1 Year Annual Desktop Licence: Buy
- Lexi-Portal 1 Month
Leximancer

Leximancer is computer software that conducts quantitative content analysis using a machine learning technique. It learns what the main concepts are in a text and how they relate to each other. It conducts a thematic analysis and a relational analysis of the textual data. Leximancer provides word frequency counts and co-occurrence counts of concepts present in the tweets. It is:

[A] Method for transforming lexical co-occurrence information from natural language into semantic patterns in an unsupervised manner. It employs two stages of co-occurrence information extraction—semantic and relational—using a different algorithm for each stage. The algorithms used are statistical, but they employ nonlinear dynamics and machine learning.

Smith and Humphreys, p. 26

Once a concept has been identified by the machine learning process, Leximancer then creates a thesaurus of words that are associated with that concept giving the ‘concept its semantic or definitional content’.
1,040,000 #BigData Tweets Analysed using Leximancer
Top Named Individuals

• @authorakansha - I'm 21 & I spent much of my time on Social Media & Photography. I'm also the author of a travel guide series

• @roymorganonline - Recognised leader in Social and Market Research, Melbourne, Australia - roymorgan.com

• @Kirkdborne - The Principal Data Scientist at @BoozAllen, PhD Astrophysicist, Data Science, Top Big Data Influencer. Ex-Professor http://rocketdatascience.org/ - Booz Allen Hamilton

• @Robertherjavec - Dad. Founder of Herjavec Group. Shark on ABC's Shark Tank. Author of You Don't Have to Be a Shark: Creating Your Own Success. Wherever I need to be. RobertHerjavec.com

• @willierandle - Follow Me, I'll Follow You! Tweet on! El Paso, TX

• @bobehayes - B.O.B. is Chief Research Officer @AnalyticsWeek. PhD in industrial-organizational psychology. Interests in #custexp #bigdata #statistics #analytics Seattle, WA • businessoverbroadway.com
Defense and Intelligence Organizations

Advanced strategy and technology solutions for our nation's defense


From battlefield to boardroom, Booz Allen's global defense experts have the experience, perspective, and know-how to support the Department of Defense's toughest missions.

Since 1914, when founder Edwin Booz helped the Navy gear up for World War II, our focus on national defense has never wavered. Today, as our military services take on new missions, adopt innovative technologies, tackle acquisition and budgeting challenges, and address warfighters' medical needs, our experts are helping the Department of Defense be ready for what’s next.

Use of the Department of Defense imagery does not constitute or imply endorsement.

---

Services
Management and Technology Consulting

Revenue
$5.48 billion (2014)

Net income
$239.955 million (FY 2014)

Number of employees
22,000 (2014)

Website
www.boozallen.com
Edward Snowden was employed by Booz Allen Hamilton and the National Security Agency.

On Twitter the key influencer around the term #bigdata is a contractor who supplies staff to the NSA.
Data Mining, Social Network Analysis and Content Analysis

• What has been presented so far are a few of the techniques and tools of data mining and analytics – with machine learning and automation in Leximancer.

• Such insights that are born from the data and the application of algorithms need to be validated in the light of informed understanding of the ‘never raw data’ position especially for matters related to health.

• The existence of this ‘data’ is the result of a long chain of requirements, goals and a shift in the wider political economy.

• The ‘insights’ are at the macro level – devoid of context and therefore an immediate sense of meaning.
Calculated Data Patients/Scientists

• Data Science generates crude quantitative knowledge, or “calculated publics” (Gillespie, 2014).

• In order to interpret the visualisations requires human perception.

• Big data and social media analytics can not answer questions about data scientists, or patients on their own.

• Big data and social media analytics requires wider knowledge of context and debates surrounding the topic at hand.

http://ethnographymatters.net/blog/2016/04/12/algorithmic-intelligence-reconstructing-citizenship-through-digital-methods/
Defined, managed, and governed

• Big/Social data does not represent what we think they do.
• Nonetheless it does represent something, and this something is certainly something worthy of our consideration.
• “That a census or a social survey is a snapshot of the way our societies are regulated is rarely remarked on and instead emphasis is given to the presumed objectivity of the categories and their data. This is the ideology of the small data era in action – the claim that it is science and not society that we are seeing through such instruments.”
• Even if the referent of population data is not the population itself, we are still dealing with reference and meaning; we are glimpsing not a population in its totality, but the various ways in which that population is defined, managed, and governed.
In Order to Interpret Data – Context and Critique are Crucial

• There are those within the Data Science discipline who are prepared to acknowledge the utility of human interpretation of data over algorithmic accounts.

• The Principal Data Scientist at @BoozAllen, (PhD Astrophysicist, Kirk Borne) recently stated that only using computer algorithms for visualisation...

“(...) can miss salient (explanatory) features of the data [therefore] a data analytics approach that combines the best of both worlds (machine algorithms and human perception) will enable efficient and effective exploration of large high-dimensional data”.

http://rocketdatascience.org/?p=567
• Unsupervised concept and theme modelling is denaturalising and unfamiliar – but crucially they are not objective, unbiased or neutral.

• They are fictions.

• The modelling algorithm knows nothing about letters, nothing about narrative form, nothing about health.

• All it can tell us is that 1) this string of tokens (in our case, words) co-occur together more than we would expect, all things being equal, and 2) some particular documents (in this case, Tweets) are composed of a certain number of tokens (words) with a relatively high probability of belonging to this topic.
• The algorithm’s lack of knowledge of semantic meaning, and particularly its lack of knowledge of the social media as a form or genre, lets it point us to a very different model of the social.

• Such ‘Reading Machines’ are engaged in datafication of the social.
Datafication

• Datafication as defined by Kenneth Cukier and Viktor Mayor-Schoenberger (2013) refers to: ‘the ability to render into data many aspects of the world that have not been quantified before’.

• This is importantly ‘not the same as digitization, which takes analog content – books, films, photographs -and converts it into digital information, a sequence of ones and zeros that computers can read.

• Datafication is a far broader activity: taking all aspects of life and turn them into data’ (Cukier & Mayor-Schoenberger, 2013). It includes ‘behavioural metadata, such as those automatically derived from smartphones, like time stamps and GPS-inferred locations’ and may be used for a range of purposes ranging from surveillance to citizen empowerment (Kennedy, Poell and van Dijck 2015).
Data is NOT Objective NOT Incontrovertible

• The concern with the notion of datafication is that as it attempts to describe a certain state of affairs, as it occurs in one moment, it also flattens human experience, in a way that ethnography always defies, by acknowledging and insisting that whatever we label as ‘data’ is ‘rich’ and ‘lively’, rather than fixed, frozen or representing something ‘true’ about the world.
• Data, particularly that which is derived from huge conglomerate sources, is becoming increasingly a material or source for driving questions for and informing design practice. This is worrying in many ways - metrics derived through Big Data always represent a partial and non-representative sample (Baym, 2013) and thus do not accurately or adequately represent how people engage with and experience the world.

• While some might fall back on the positivist argument that we simply need to improve our measures, people –as everyday designers– will intentionally or unintentionally ensure that data is incomplete, dispersed and unfinished.
• There are growing movements towards data as fiction – embodied within data is an ideology.

• More recently for example the unfitbits online initiative ([http://www.unfitbits.com/](http://www.unfitbits.com/))

• Yet what is perhaps more worrying is that designers, developers, and policy makers will continue to take Big/Social Data at face value, as an object or representation of a truth that can be extracted from and that reflects the social.

• We are glimpsing the *various ways in which we are defined, managed, and governed.*