Coding the Twitter Sphere: Humans and Machines Learning Together

Dr. Stuart Shulman
@stuartwshulman
stu@texifter.com
Acknowledgements

The National Science Foundation
Mark J. Hoy
Conflict of Interest Disclosure

I am the sole manager of Texifter

We sell DiscoverText licenses

We sell Gnip data licenses
A Master Metaphor: Sifter
Welcome to the Coding Analysis Toolkit (CAT)

CAT is a free service of the Qualitative Data Analysis Program (QDAP), and hosted by the University Center for Social and Urban Research, at the University of Pittsburgh, and QDAP-UMass, in the College of Social and Behavioral Sciences, at the University of Massachusetts Amherst. CAT was the 2008 winner of the "Best Research Software" award from the organized section on Information Technology & Politics in the American Political Science Association.

For the CAT Quick Start Guide, you can view the PDF file CAT Quickstart Guide or watch the CAT Tutorial - February 23, 2009

May 5, 2010 - CAT is now an open source project! You can host your own version of CAT from the project source code at: http://sourceforge.net/projects/catoolkit/

Read about the latest innovations in text analytics on the Textifier blog or via a specially curated @scoopit page

CAT Statistics
There are currently 5,936 primary CAT accounts and 1,014 sub-accounts. CAT users have uploaded 5,793 coded datasets and 8,425 raw datasets. They have coded a total of 1,482,741 items and adjudicators have made 164,177 validation choices in CAT.

What can you do in CAT?
- Efficiently code raw text data sets
- Annotate coding with shared memos
- Manage team coding permissions via the Web
- Create unlimited collaborator sub-accounts
- Assign multiple coders to specific tasks
- Easily measure inter-rater reliability
- Adjudicate valid & invalid coder decisions
- Report validity by dataset, code or coder
- Export coding in RTF, CSV or XML format
- Archive or share completed projects

What file types can CAT import?
- Plain text
- HTML
- CAT XML
- Merged ATLAS.ti coding

CAT Resources
- Raw Data Preparation Guide
- ATLAS.ti Upload Preparation

Have you tried DiscoverText?
Featuring the Facebook Graph & Twitter APIs
Three Primary Tasks in CAT

What can you do in CAT?

- Efficiently code raw text data sets
- Annotate coding with shared memos
- Manage team coding permissions via the Web
- Create unlimited collaborator sub-accounts
- Assign multiple coders to specific tasks
- Easily measure inter-rater reliability
- Adjudicate valid & invalid coder decisions
- Report validity by dataset, code or coder
- Export coding in RTF, CSV or XML format
- Archive or share completed projects
Classification of Text

A 2500 year-old problem

Plato argued it would be frustrating

It still is...
Volume is a problem for scholars
Coders are expensive
Groups struggle to accurately label text at scale
Validation of both humans and machines is “essential”
Some models are easier to validate than others
All models are wrong
Automated models enhance/amplify, but don’t replace humans
There is no one right way to do this
“Validate, validate, validate”
“What should be avoided then, is the blind use of any method without a validation step.”
<table>
<thead>
<tr>
<th>Coder:</th>
<th># Datasets</th>
<th># Annotations</th>
<th># Validations</th>
<th>% Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>31</td>
<td>15480</td>
<td>313</td>
<td>96.49%</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>7676</td>
<td>351</td>
<td>82.62%</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>18300</td>
<td>2790</td>
<td>93.26%</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>15283</td>
<td>276</td>
<td>42.75%</td>
</tr>
<tr>
<td></td>
<td>109</td>
<td>84793</td>
<td>7202</td>
<td>91.53%</td>
</tr>
</tbody>
</table>

(Patent Pending)
Three Important Books

"Google is not just a company, it is an entirely new way of thinking about understanding who we are and what we want. Jarvis has done something really important: extend that approach to business and culture, revealing just how revolutionary it is." – Chris Anderson, author of The Long Tail

What Would Google Do?

James Gleick

The Information

A History, A Theory, A Flood

David Weinberger

Everything Is Miscellaneous

The Power of the New Digital Disorder

“Perfectly placed to tell us what’s really new about the second-generation Web.”
—Los Angeles Times
One Particularly Important Idea

Crowdsourcing brings widely distributed wisdom to process of text analysis

“This is really the biggest paradigm shift in innovation since the Industrial Revolution”
- MIT professor Eric von Hippel, specialist in innovation management
Five Pillars of Text Analytics

Search
Filter
Code
Cluster
Classify

You can execute all five using DT
Pillar #1: Search

... that shutting down traffic was his “favorite thing about being Governor” - http://t
... other affected E.Coast traffic. Very bad. Obama HHS & IRS abuse of power affects millions natio
... not called over GWB traffic - http://t.co/0FMHGnrBwL http://t.c
... erage of Christie’s Traffic Scandal Than in Last Six Months on IRS http://t.co/FtvB62n7G
... er affected E.Coast traffic. Very bad. Obama HHS & IRS abuse of power affects millions natio
... ee them languish in traffic, and hear the lamentations of their
... hristie ordered the traffic jam. How do u feel?
... KE ON GOV CHRISTIE? TRAFFIC JAM APOLOGY? CHRISTIE IS TOOOOOOO FAT!! FRACK THE FAT’ CABOT OIL& GA
... logan: Think of the traffic studies I can do with the presidential motorcade. #bridgegat
... t knowing about the traffic is like Obama not knowing about Benghazi is he giving O a pa
... ave been broken' in traffic scandal: http://t.co/KZmG9F8pp
... ave been broken' in traffic scandal: http://t.co/c3kMAWLyg
... ave been broken' in traffic scandal: http://t.co/SuSBcnRJ3
... ave been broken' in traffic scandal: http://t.co/gfoT5WPC1
... not called over GWB traffic http://t.co/wSHmVwymZS #bridgegat
... of blocking so much traffic it’s Christie. (Pause) He's a very large man
... not called over GWB traffic http://t.co/wSHmVwymZS #bridgegat
Search for Negative Cases

Agatha Christie
Writer

Dame Agatha Mary Clarissa Christie, DBE was an English crime writer of novels, short stories, and plays. [Wikipedia]

**Born:** September 15, 1890, Torquay, United Kingdom

**Died:** January 12, 1976, Wallingford, Oxfordshire, United Kingdom
Defined Search (Multi-term)
Pillar #2: Filters

Search and Browse Dataset

Relevant Tweets? *(Dataset)*
Showing 1 to 62 of 62 total Filters applied

Add to Bucket: New | Existing | Selected

Advanced filters
Search query:
Defined search: -- use pre-defined search -
Filter by date: to
Filter by meta:
Filter by annotations:
Filter by coding:
Filter by classification:
Set Classification Bounds: set classification boundary filter
Selected filters:
No, not Christie Winning
(items not coded)
Another Common Filter

<table>
<thead>
<tr>
<th>Advanced filters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Search query:</td>
<td></td>
</tr>
<tr>
<td>Defined search:</td>
<td>-- use pre-defined search -</td>
</tr>
<tr>
<td>Filter by date:</td>
<td></td>
</tr>
<tr>
<td>Filter by meta:</td>
<td>created_at: (Date)   =</td>
</tr>
<tr>
<td>Filter by annotations:</td>
<td>Has File Annotations</td>
</tr>
<tr>
<td>Filter by coding:</td>
<td>(all coded items)</td>
</tr>
<tr>
<td>Selected filters:</td>
<td>from_user:           Contains</td>
</tr>
<tr>
<td></td>
<td>(items not coded)</td>
</tr>
</tbody>
</table>
### Top values for: favourites_count:

<table>
<thead>
<tr>
<th>Meta Value</th>
<th>Total</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>146363</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>45411</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>44173</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>43412</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>43411</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>41698</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>39859</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>39195</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>36395</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>36394</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 458 total
Pillar #3: Human Coding

Annotation enhances your analysis by applying human interpretation to machine results
Keystroke Coding is Fast
Coding Off a List is Faster

Search and Browse Dataset

Relevant Tweets v2 (Dataset)
Showing 1 to 100 of 793 total

- RT @warriorwoman9 RT @LindaP09
  http://t.co/0uwidgG4u
- ...ts New Theory About
http://t.co/loBpbUQ5k
- ...T @lordxm2k: The "Chris Christie" Code: He Used Sandy Relief Funds To Make Himself Look Good & The Bridge Lane Closures To Exact Revenge
- Rove: Bridge scandal shows Chris Christie is 'what we want' in a president http://t.co/KKompActRI lol #p2 #tcot #uniteblue #pjnet #lnyhb
- ...ts New Theory About Chris Christie's Bridge GionnyScandalChris Christie, Mark Sokolich, Chris...
http://t.co/UScODPigF
- ...s expected for Gov. Chris Christie aides over bridge scandal: http://t.co/mMhQEMiALJ
- The "Chris Christie" Code: He Used Sandy Relief Funds To Make Himself Look Good & The Bridge Lane Closures To Exact Revenge On Democrats
- ...O to Rightbloggers, Chris Christie Finally Earns Respect With His Bridge to Benghazi http://t.co/sh4N0HnaE5 via @villagevoic
- Karl Rove: Chris Christie's Bridge Scandal Response Gives 'Street Cred' with Tea Party http://t.co/xWkqdPNorh via @BreitbartNews BS KAR
- Say What?: Chris Christie Bridge Scandal - KGAN-TV CBS 2 Iowa - Top Stories: http://t.co/WNfxKDiyl
- ...s expected for Gov. Chris Christie aides over bridge scandal: http://t.co/MBG9kTA0v

Showing 1 to 100 of 793 total
Data Cleaning is Fundamental

This unit is coded with: *No, off topic*
Pillar #4: Clustering

Archive name: Christie - Twitter API

Near-duplicates (1839 clusters)

<table>
<thead>
<tr>
<th>Similar Clusters</th>
<th>Cluster Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>(214)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(157)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(130)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(94)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(78)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(77)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(66)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(65)</td>
<td>create bucket</td>
</tr>
<tr>
<td>(60)</td>
<td>create bucket</td>
</tr>
</tbody>
</table>
Making federal case against Christie on lane closures tough: expert - New York Daily News: New http://t.co/mCMVRriPwa
Latent Dirichlet Allocation (LDA) Topic Models
LDA on the Christie Data

Topic 1: christie, sandy, christies, funds, relief, feds, investigating, daily, gov, feminized
Topic 2: with, daniel, didnt, after, murder, time, agatha, death, former, mayor
Topic 3: bridge, about, traffic, more, scandal, chris, nj, some, just, says
Topic 4: like, gop, bridgegate, what, 2016, know, now, will, bully, dont
Topic 5: obama, benghazi, impeachment, dem, have, probe, lawmaker, floats, possibility, gwob
Topic 6: jersey, over, stages, still, aides, grief, bogus, hes, news, subpoenas
Topic 7: rove, closures, karl, york, while, federal, party, tea, governor, president
Topic 8: irs, political, been, show, republicans, media, get, laws, word, scandals
Pillar #5: Machine-Learning

What Humans Do Best

What Computers Do Best

Humans and machines learning together

Active Learning
Create a Dataset to Code

Any archive or bucket

Use the random sampling tool

Standard: All coders get all items

Triage: Coders get next uncoded item
Select from Three Coding Styles

Default: Mutually Exclusive Codes

Option 1: Non-Mutually Exclusive Codes

Option 2: User-Defined Codes (Grounded Theory)
Assign Peers to Code a Dataset

How many coders?

How many items need to be coded?

How many test or training sets?

There are no cookbook answers
Look at Inter-Rater Reliability

Highly reliable coding (easy tasks)

Unreliable coding (interesting tasks)

If humans can’t, neither can machines

Some tasks better suited for machines
Adjudication: The Secret Sauce

Expert review or consensus process

Invalidated false positives

Identify strong and weak coders

Exclude false positives from training sets
Adjudicate Dataset

Dataset Details > Validate Dataset

Dataset: Fear Arousal >95 v1

Code: **Fear Rejection**  (1) Valid  (2) Skip to next  (3) Not valid

40.00% of users coded this as Fear Rejection

- Soooo Tired of Dat Commercial of Tha Woman Name Terry w| Tha Hole n Her Throat & Da Blonde Wiq
#NoOffense

Coder choices

<table>
<thead>
<tr>
<th>Coder</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shulman, Stu</td>
<td>Fear Rejection</td>
</tr>
<tr>
<td>Pruitt, Katie</td>
<td>Fear Arousal</td>
</tr>
<tr>
<td>Apostoleris, Lucas</td>
<td>Fear Arousal</td>
</tr>
<tr>
<td>Liew, Jasy</td>
<td>Fear Neutral</td>
</tr>
<tr>
<td>Eanes, Ryan</td>
<td>Fear Rejection</td>
</tr>
</tbody>
</table>
Holy shit that commercial of that lady Terri who used to be a smoker is freaky
Use Classification Scores as Filters

Iteration plays a critical role
Train, classify, filter
Repeat until the model is trusted
Each round weeds out false positives
Classifier Histograms: More Filtering

Classification Boundary Filter

Classification Histogram

Histogram Range: 0% to 100%

select all  clear all  apply filter
Asifter

Search and retrieve data from the complete history of Twitter

http://sifter.texifter.com
Rules & Filtering

Before the PowerTrack API will deliver activities to you through the streaming connection, it must be configured with a set of rules. These rules are applied across the firehose of the data source, and with those activities that match being sent through your API connection.

Put simply, PowerTrack rules allow you to build a list of requirements for the data that will be sent through your stream. For example:

*I want activities where the text mentions the keyword “gnip”*

Or, for a more realistic example:

*I want activities which meet the following requirements*

- contain any of these keywords: gnip, data, social
- and contain any of these keywords: love, hate, like, dislike
- and which contain a link
- but which don’t contain any of these keywords: ping, gnop

The above requirements can be translated into a PowerTrack rule:

```
(gnip OR data OR social) (love OR hate OR like OR dislike) has:links -(ping OR gnop)
```

Operators

In the example above, the keywords and ‘has:links’ are Operators. Operators tell PowerTrack what types of activities to deliver. For example, the keyword operator ‘gnip’ above tells PowerTrack that an activity must contain a tokenized keyword in the text, while the has:links operator adds a requirement that the activity contain a URL in its body.

Operators may be either positive or negative.
Thanks for Listening

Dr. Stuart Shulman
@stuartwshulman
stu@texifter.com
discovertext.com
sifter.texifter.com