"...a wealth of information creates a poverty of attention."
- Herbert Simon, 1971

Automated Text Coding: Humans and Machines Learning Together

Dr. Stuart W. Shulman
Founder & CEO, Texifter
@stuartwshulman
Text Classification

A 2500 year-old problem

Plato argued it would be frustrating. It still is.
THE COTTON ASSOCIATION vs. THE COTTON EXCHANGE

Mistake to Suppose That Either “Long” Buying or “Short” Selling Depresses Prices—Even the Bad Grading Methods of the Cotton Exchange Does Not Keep Prices Down

By Thomas Gidson

The Southern Cotton Association has declared bitter war on the New York Cotton Exchange. The representatives of the Southern Association claim that prices have been kept down for years, largely on account of the acts of speculators, and that the time has arrived to call a halt. At the annual meeting of the Southern Cotton Association, held at Birmingham, Ala., on Jan. 17th, Mr. M. L. Johnson, president of the Georgia division, expressed himself on the subject thus:

“For forty years we have paid tribute to Wall Street gamblers.”

Mr. Harris Jordan, president of the Association, on the same occasion said:

“This Association has no desire to make war upon any Cotton Exchange which issues an honest, legitimate contract, but all forms of gambling and speculation we will continue to combat until relief for the cotton growers has been fully obtained.”

Verily, if Mr. Jordan is to accomplish these things he has his work cut out for him.

Does Speculation Depress Prices?

But it is with the speculative phase of the matter that this article has to deal.

In all its tirades against the Cotton Exchange, the representatives of the Southern Cotton Association have taken the ground that speculative movements depress prices. This is exactly the reverse of the truth. Almost all great speculative campaigns have been for high prices. In cotton—the Prize-McCormick campaign, the Sully campaign and dozens of minor deals stand out in bold relief; just as the operations of Old Hutch, Harper, Leiter, Coster-Martia and Phillips are landmarks in cereal speculation. The high prices established in such movements offer an

Rates and Terms for Credit
Farm Profitability
Cost of Living
Soil Fertility
Education
Relations between Classes
Fall 1999
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.”
Welcome to the Coding Analysis Toolkit (CAT)

CAT is a free service originally by the Qualitative Data Analysis Program (QDAP), and was 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 here:
CAT Quickstart Guide

To view a tutorial on using CAT, click here:
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/

CAT Statistics
There are currently 10,073 primary CAT accounts and 1,410 sub-accounts. CAT users have uploaded 7,971 coded datasets and 13,252 raw datasets. They have coded a total of 2,135,687 items and adjudicators have made 178,180 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
- Merging HUs in ATLAS.ti

Have you tried DiscoverText?
Featuring the Facebook Graph & Twitter APIs

If you like CAT, you’ll love DiscoverText. DiscoverText is a cloud-based, collaborative text analytics solution. Generate valuable insights about customers, products, employees, news, citizens, and more. Sign up for a 30 day free trial.

© 2007-2016 - CAT is maintained by Texifter, LLC and powered by Microsoft ASP.NET.
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
Higher efficiency and less wrist pain

We have implemented an efficient workflow to help you avoid those aching joints.

Avoid Tennis Elbow

Items load to the screen and the coder hits the keystroke
**Keystroke Human Coding**

Human coding can be distributed to individuals, groups & crowds.

---

<table>
<thead>
<tr>
<th>(1) News or Current Events</th>
<th>(2) Sports</th>
<th>(3) Food or Restaurants</th>
<th>(5) Other</th>
</tr>
</thead>
</table>

**Data**

Europafoot
@Europafootcom

Mercato - Lyon : Un grand d'Angleterre rêve de Lacazette
- europafoot.com/mercato-lyon-u... dans #Football

7:52 PM - 8 Apr 2015

**Codes**
Computer Science & NSF Influence: Measure Everything

How fast?

How reliable?

How accurate?

How fast?
Annotator Speed

<table>
<thead>
<tr>
<th>Coder</th>
<th>Units Coded</th>
<th>Avg. Coding Time</th>
<th>Total Coding Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (100.00%)</td>
<td>6.52s</td>
<td>00:21:44</td>
<td></td>
</tr>
<tr>
<td>200 (100.00%)</td>
<td>4.65s</td>
<td>00:15:30</td>
<td></td>
</tr>
<tr>
<td>200 (100.00%)</td>
<td>10.32s</td>
<td>00:34:24</td>
<td></td>
</tr>
<tr>
<td>200 (100.00%)</td>
<td>4.30s</td>
<td>00:14:19</td>
<td></td>
</tr>
<tr>
<td>200 (100.00%)</td>
<td>4.30s</td>
<td>00:14:19</td>
<td></td>
</tr>
<tr>
<td>200 (100.00%)</td>
<td>7.27s</td>
<td>00:24:14</td>
<td></td>
</tr>
</tbody>
</table>
Interrater Reliability: A Critical Measurement

<table>
<thead>
<tr>
<th>Code</th>
<th>Coder 5981</th>
<th>Coder 9953</th>
<th>Coder 9955</th>
<th>Coder 2668</th>
<th>Coder 2279</th>
<th>Coder 4341</th>
<th>Exact Match</th>
<th>Partial Match</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH-17</td>
<td>11</td>
<td>13</td>
<td>27</td>
<td>20</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>12</td>
<td>0.69</td>
</tr>
<tr>
<td>Not MH-17</td>
<td>189</td>
<td>187</td>
<td>173</td>
<td>180</td>
<td>192</td>
<td>189</td>
<td>172</td>
<td>17</td>
<td>0.96</td>
</tr>
<tr>
<td>Totals</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>179</td>
<td>29</td>
<td>0.94</td>
</tr>
</tbody>
</table>
MH 17 事后第一天，美国就宣布其一套雷达系统看到了一枚地空导弹从乌克兰境内飞向马航，还有另一套看到了马航被击中的热信号，如此精确和反应迅速的雷达系统……在几个月前，为啥却看不到马航370……一架在空中飞了8个小时……比导弹大那么多倍的777……去哪了？
CoderRank for enhanced machine-learning is our key innovation

Dataset: MH17 Test 1

Total Valid Answers: 1010 / 1209 (83.54%)

Validations by Coder:
- Coder 3510: 4 / 4 (100.00%)
- Coder 9817: 197 / 200 (98.50%)
- Coder 5611: 196 / 200 (98.00%)
- Coder 0043: 195 / 200 (97.50%)
- Coder 9783: 169 / 200 (84.50%)
- Coder 5981: 124 / 200 (62.00%)
- Coder 2279: 122 / 200 (61.00%)
- Coder 6549: 3 / 5 (60.00%)

Validations by Code:
- MH-17: 672 / 681 (98.68%)
- Not MH-17: 338 / 528 (64.02%)

Dataset: MH17 Test 2

Total Valid Answers: 734 / 800 (91.75%)

Validations by Coder:
- Coder 5611: 197 / 200 (98.50%)
- Coder 9783: 184 / 200 (92.00%)
- Coder 0043: 179 / 200 (89.50%)
- Coder 5981: 174 / 200 (87.00%)

Validations by Code:
- MH-17: 564 / 567 (99.47%)
- Not MH-17: 170 / 233 (72.96%)

Patent issued March 1, 2016
CoderRank is to text analytics what PageRank was to search. Just as Google said not all web pages are created equal, Texifter argues that not all humans are created equal. When training machines, it is best to rely most on the humans most likely to create a valid observation. We proposed a unique way to rank humans on trust and knowledge vectors.
The Five Pillars of Text Analytics

Search
Filtering
De-duplication and Clustering
Human Coding
Machine-Learning
Pillar #1: Search
Pillar #2: Filters

TopMeta Discovery

Top values for: postsinthread:

<table>
<thead>
<tr>
<th>Meta Value</th>
<th>Total</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>230212</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>230062</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>230061</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>42392</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>22412</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21825</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21360</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21359</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21358</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21357</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 657 total
Pillar #3: Deduplication & clustering

![Active Clustering](image)

<table>
<thead>
<tr>
<th>Group Id</th>
<th>Items</th>
<th>Snippet</th>
</tr>
</thead>
</table>
| Group 1  | 110   | RT@renfanzi: 俄罗斯电视台（今日俄罗斯）女记者沙拉•佛丝（Sara Firth）昨日在推特宣布辞职。她
         |       | 被称作“克里姆林宫的女炸弹记者”。推特用户表示，她将
         |       | 暂停推特活动。推特用户表示，她将
         |       | 暂停推特活动。推特用户表示，她将
         |       | 暂停推特活动。推特用户表示，她将
         |       | 暂停推特活动。推特用户表示，她将
         | Group 2 | 62    | RT@shifeike: 马航乌克兰空难的热
         |       |       | 点不
         |       |       | 有
         |       |       | 比较
         |       |       | 比较
         |       |       | 比较
         |       |       | 比较
         | Group 3 | 55    | RT@renfanzi: 乌克兰亲俄叛军找到了被击落的
         |       |       | 民航飞机的第一个黑匣子。第一个反
         |       |       | 应是往俄罗斯送，幸亏老毛子反应
         |       |       | 还算比较快，从乌克兰西部
         |       |       | 进行
         |       |       | 比较
         | Group 4 | 42    | RT@duanlidream: 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         |       |       | 俄罗斯电视台
         | Group 5 | 31    | RT@kirk 1031: 美国驻乌克兰大使馆对MH17被击落事件的评估："反对派用SA11导弹击
         |       |       | 落
         |       |       | 的
         |       |       | 的
         |       |       | 的
         |       |       | 的
         |       |       | 的
         | Group 6 | 29    | RT@uponsnow: CNN的慢
         |       |       | 播
         |       |       | 播
         |       |       | 播
         |       |       | 播
         |       |       | 播
         | Group 7 | 22    | RT@huangboma: 万科的
         |       |       | 经</p>

---

207 total groups and 1686 single items

Items per page: 20
Page: 1 2 3 4 5 6 7 8 9 10 11
Pillar #4: Human coding (a.k.a. labeling or tagging)
Pillar #5: Machine-learning
Active Learning engines and human coding tools combine...

what humans do best... with what computers do best

Humans and machines learning together

It is always good to keep humans “in-the-loop”
Word sense disambiguation (relevance)
Word sense disambiguation (relevance)
Word sense disambiguation (relevance)
Word sense disambiguation (relevance)

Yes
No
No
Human coding can be converted into machine classifiers

Accumulated human coding becomes training data via machine-learning
Crowdsourcing accelerates the insight generation process through machine-learning

Distributed for synchronous & asynchronous collaboration
Thank–you for listening!

@stuartwshulman
stu@texifter.com