National Security & Intelligence
Applications of Text Analytics

Patrick Lam

Lead Data Scientist, Thresher
Visiting Fellow, Harvard IQSS

March 26, 2015
What is text analytics?
"a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation"

- Wikipedia
Knowledge
What are some common text analysis methods?
Natural Language Processing (NLP)

Enable computers to understand human text input.
Machine Learning

Grouping, classifying, predicting
Applications
Who Authored the Federalist Papers?

(Mosteller and Wallace, 1963)
43 papers
Hamilton

14 papers
Madison

5 papers
Jay

12 papers
H, M, or J?
Model the usage of high-frequency function words for each author.

also, and, by, of, on, there, ...
43 papers
Hamilton

14 papers
Madison

5 papers
Jay

12 papers
Madison
Authorship attribution using text analysis has a large literature and can be useful for national security & intelligence purposes.
Harvard IQSS
Institute for Quantitative Social Science

iq.harvard.edu
Four recent projects using text analysis by Harvard IQSS affiliates with relevance to national security & intelligence
1. Reverse-Engineering Censorship In China

(King, Pan, and Roberts, 2013 and 2014)
What Gets Censored In China?

Monitor ~1,400 Chinese social media sites over 6 months across 85 content areas.

Download posts the instant they appear.

Revisit each post later to check if it was censored.

Analyze with new methods of text analysis.

Experiment with writing social media posts.
Censorship program targets collective action rather than criticism of the government.
2. Jihadi Radicalization of Muslim Clerics

(Nielsen, 2014)
Why Do Some Clerics Preach Jihad While Others Do Not?

Download writings of a sample of Muslim clerics.

Use machine learning methods to score clerics on level of jihad by comparing writings to known Jihadi texts.

Analyze along with other data on clerics.
Clerics with weak educational networks and connections often use Jihadi ideology to appeal to lay audiences and further their careers.
3. Predicting Crowd Behavior

(Kallus, 2014)
Can We Predict Major Protests?

Scan over 300,000 web sources (news, blogs, forums, Twitter) in 7 languages for mentions of past, current, or future events.

Extract type of event, entities involved, and timeframe using NLP methods.

Predict on each day whether a significant protest will occur over the next three days using machine learning methods.
Massive online public discourse data has the potential to predict crowd behavior using text analysis methods.
4. Anti-Americanism in the Middle East

(Jamal, Keohane, Romney, and Tingley, 2015)
Anti-Americanism on Twitter

Download Arabic Twitter posts by keywords.

Consider discourse about US in general and in reaction to specific events.

Use text analysis methods to classify proportion of posts in specified categories.
Anti-Americanism in the Middle East is directed toward the impingement of the US on other countries rather than toward American society.
All applications of text analysis have a common starting point.
How do we define our set of texts that we want to analyze?
How do we retrieve the relevant texts from the vast set of texts available?
How do we *follow* the relevant conversations as they evolve?
Define.
Retrieve.
Follow.
In some cases, the relevant set of texts is well-defined, static, and easily accessible.

Federalist Papers, complete works of Shakespeare
In most cases, we have to constantly search and retrieve the relevant texts, often using *keywords and Boolean searches*. Twitter, news articles, blogs and forums
Our analyses are only as good as our keywords!
What is our current most commonly used technology for defining the relevant keywords to retrieve our texts?
Example: Think of keywords you would use to follow the Twitter conversation around the Boston Marathon Bombings.
Keywords about the event

#bostonbombings, explosion, terrorism, attack, ...
Keywords about the suspects

suspect, tsarnaev, dzhokhar, tamerlan, ...
Keywords about the victims

innocent, victim, collier, ...
Keywords about the reaction

tragedy, prayers, #prayforboston, …
Keywords about the politics

obama, #tcot, #benghazi, ...
We ran a similar experiment with 43 Harvard undergrads.
59% of the words were suggested by only 1 out of 43 undergrads.
Median number of words per respondent was 7.
Humans are good at recalling a small list of good keywords and recognizing a good keyword when they see it.
Humans are bad at recalling a long list of keywords that capture different ways of representing a concept.
Some Existing Options for Automated Keyword Discovery
1. Mine search queries

Google Adwords
2. Thesaurus methods

reference books, WordNet
3. Co-occurrence methods
Enter Thresher.

(Based on King, Lam, and Roberts, 2014)
Thresher keeps humans in the loop and helps them find more and better words faster.
The Thresher Algorithm

Reference set $R$: texts about concept of interest

Search set $S$: broad set of texts

Target set $T$: texts in $S$ about the same concept as in $R$

Goal: Estimate $T$ and find keywords that define $T$. 
Thresher Applications
The Boston Marathon Bombings on Twitter

R: #bostonbombings
S: boston
Thresher separates relevant words from irrelevant words.
<table>
<thead>
<tr>
<th>Target set words</th>
<th>Non-target set words</th>
</tr>
</thead>
<tbody>
<tr>
<td>suspect, bomb, police, people, fbi, tsarnaev, arrest, die, terror, attack, kill, obama, custody, prayer, dzhokhar, god, false, #prayforboston, #tcot, #bostonmarathon, picture, identify, russia, #watertown, tamelan, islam, jihad,...</td>
<td>game, red sox, bruins, celtics, back, tonight, fan, #mlb, night, chicago, new york, garnett, fenway, rondo, #job, playoff, yankees, blackhawks, stanley, pizza, #nhl, draft,...</td>
</tr>
</tbody>
</table>
The Bo Xilai Scandal on Chinese blogs and forums

R: 薄熙来 (Bo Xilai)
S: 重庆 (Chongqing)
王立军 Wang Lijun
政治 government
事件 event (euphemism for the scandal)
打黑 strike corruption
犯罪 commit a crime
民主 democracy
权力 power
文革 Cultural Revolution
领导 leader
改革 reform
群众 the masses
中央中共 Central Communist Party
社会主义 socialism
唱红 sing red songs
黑社会 black society
干部 cadre
路线 party line
Writings About Suicide Bombings in Arabic

R: (martyrdom operations) from "Haqibatu'I-Mujahid" (A Mujahid's Bookbag)

S: "Pulpit of Tawhid and Jihad"
A Jihadist web library
enemy
killing
to vex or spite ("vex the infidels")
teach them
steed
fight
wronged
terrify
boy
(the story of the boy and the king, relevant to jihadis)
"And prepare against them whatever you are able of power and of steeds of war by which you may terrify the enemy of Allah and your enemy and others besides them whom you do not know [but] whom Allah knows. And whatever you spend in the cause of Allah will be fully repaid to you, and you will not be wronged."

- Quran 8:60
Can we use Thresher to define different Arabic dialects by keywords and retrieve texts from them?
<table>
<thead>
<tr>
<th>Egyptian</th>
<th>Gulf</th>
<th>Levantine</th>
<th>MSA</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ايه</td>
<td>وش</td>
<td>شو</td>
<td>ماذا</td>
<td>What</td>
</tr>
<tr>
<td>معرفش</td>
<td>مادري</td>
<td>ما يعرف</td>
<td>لا يعرف</td>
<td>Don't know</td>
</tr>
<tr>
<td>عايز</td>
<td>ببي</td>
<td>بدي</td>
<td></td>
<td>To want/I want</td>
</tr>
<tr>
<td>عاوز</td>
<td>بغي</td>
<td>بدق</td>
<td></td>
<td>To want/You want</td>
</tr>
<tr>
<td>عايزين</td>
<td>ببون</td>
<td>بدمك</td>
<td></td>
<td>To want/You want (pl.)</td>
</tr>
</tbody>
</table>
Conversations can evolve quickly in response to certain actors and we need to be able to follow them.
Evading Censors in China
自由  Freedom
目田  Eye field
(homograph)
和谐  Harmonious [Society]
河蟹  River crab
(homophone)
Evading Authorities and the Distribution of Child Pornography
PatrickLam.org
ThresherVentures.com
iq.harvard.edu