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


digg
You
Tube


.word
.pdf
.zip
.xls


Education
Information
Solution
Experience
Analysis
Learning
Knowledge

What are some common text analysis methods?

Natural Language Processing (NLP)

Enable computers to understand human text input.

Machine Learning

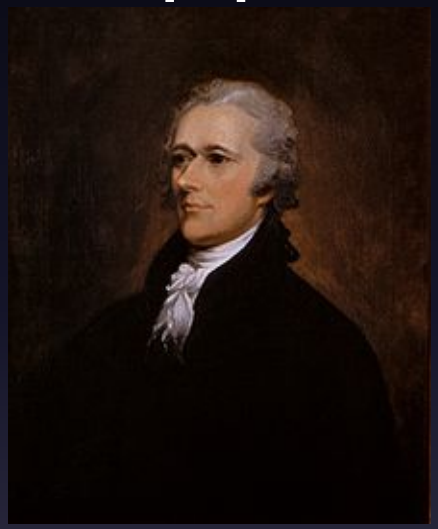
Grouping, classifying, predicting

Applications

Who Authored the Federalist Papers?

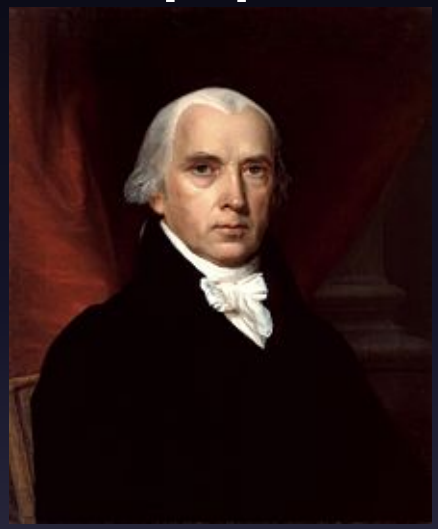
(Mosteller and Wallace, 1963)

43 papers



Hamilton

14 papers



Madison

5 papers

Jay

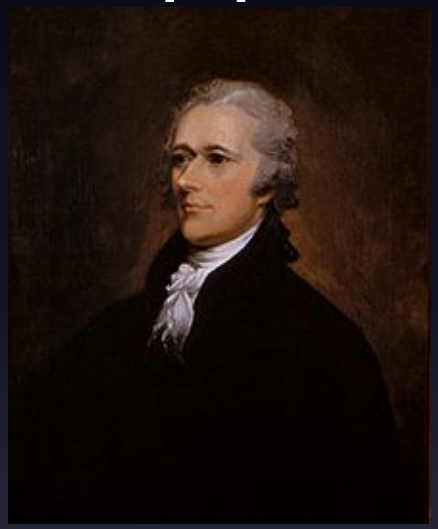
12 papers

H, M, or J?

Model the usage of high-frequency function words for each author.

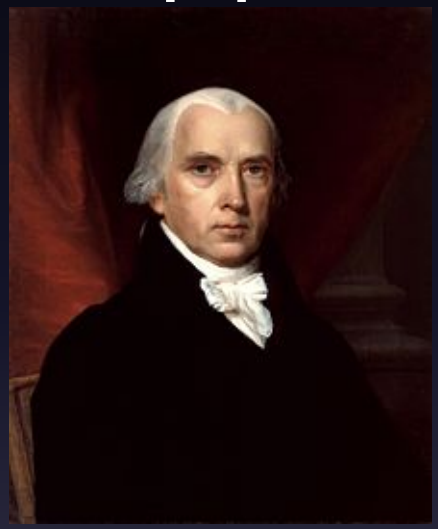
also, and, by, of, on, there, …

43 papers



Hamilton

14 papers

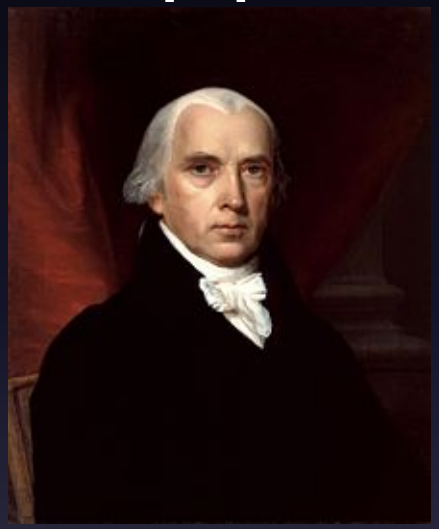


Madison

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?

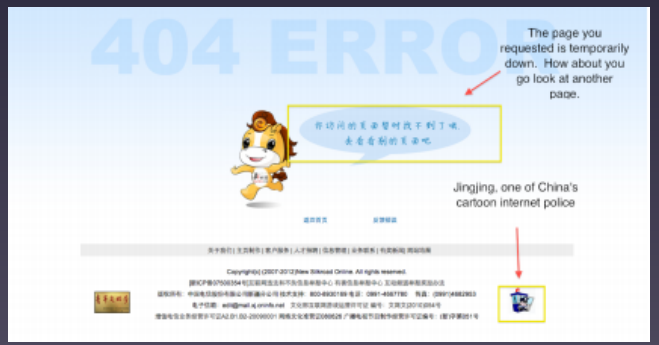
Monitor ~1,400 Chinese social media sites over 6 months across 85 content areas.

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

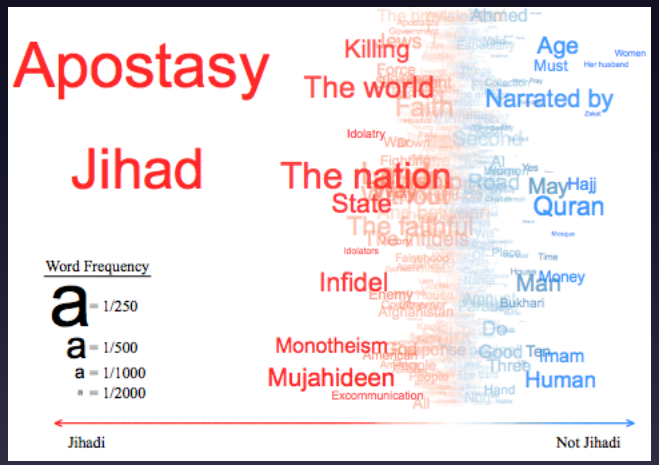
(Nielsen, 2014)

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

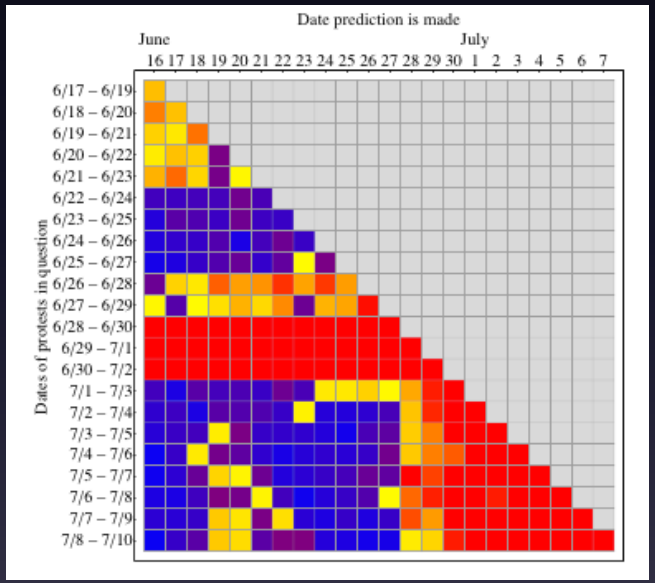
(Kallus, 2014)

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

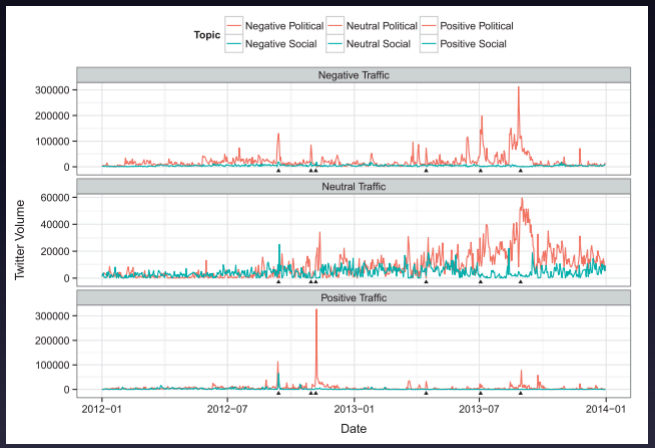
(Jamal, Keohane, Romney, and Tingley, 2015)

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.

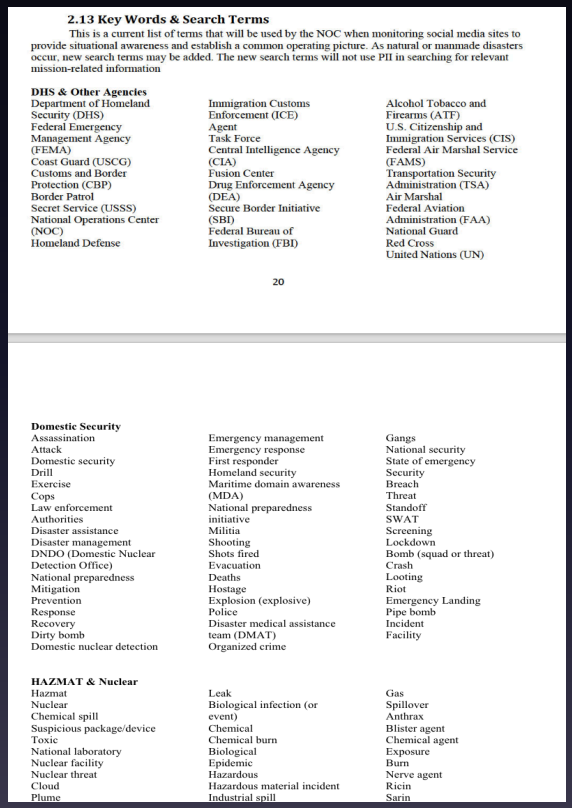
In some cases, the relevant set of texts is well-defined, static, and easily accessible.

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

DHS Analyst's Desktop Binder



2.13 Key Words & Search Terms

This is a current list of terms that will be used by the NOC when monitoring social media sites to provide situational awareness and establish a common operating picture. As natural or manmade disasters occur, new search terms may be added. The new search terms will not use PII in searching for relevant mission-related information

DHS & Other Agencies

Department of Homeland Security (DHS)
Federal Emergency Management Agency (FEMA)
Coast Guard (USCG)
Customs and Border Protection (CBP)
Border Patrol
Secret Service (USSS)
National Operations Center (NOC)
Homeland Defense
Immigration Customs Enforcement (ICE)
Agent
Task Force
Central Intelligence Agency (CIA)
Fusion Center
Drug Enforcement Agency (DEA)
Secure Border Initiative (SBI)
Federal Bureau of Investigation (FBI)
Alcohol Tobacco and Firearms (ATF)
U.S. Citizenship and Immigration Services (CIS)
Federal Air Marshal Service (FAMS)
Transportation Security Administration (TSA)
Air Marshal
Federal Aviation Administration (FAA)
National Guard
Red Cross
United Nations (UN)

Domestic Security

Assassination
Attack
Domestic security
Drill
Exercise
Cops
Law enforcement
Authorities
Disaster assistance
Disaster management
DNDO (Domestic Nuclear Detection Office)
National preparedness
Mitigation
Prevention
Response
Recovery
Dirty bomb
Domestic nuclear detection
Emergency management
Emergency response
First responder
Homeland security
Maritime domain awareness (MDA)
National preparedness initiative
Militia
Shooting
Shots fired
Evacuation
Deaths
Hostage
Explosion (explosive)
Police
Disaster medical assistance team (DMAT)
Organized crime
Gangs
National security
State of emergency
Security
Breach
Threat
Standoff
SWAT
Screening
Lockdown
Bomb (squad or threat)
Crash
Looting
Riot
Emergency Landing
Pipe bomb
Incident
Facility

HAZMAT & Nuclear

Hazmat
Nuclear
Chemical spill
Suspicious package/device
Toxic
National laboratory
Nuclear facility
Nuclear threat
Cloud
Plume
Leak
Biological infection (or event)
Chemical
Chemical burn
Biological
Epidemic
Hazardous
Hazardous material incident
Industrial spill
Gas
Spillover
Anthrax
Blister agent
Chemical agent
Exposure
Burn
Nerve agent
Ricin
Sarin

Strain
Quarantine
H1N1
Vaccine
Tamiflu
Norvo Virus
Epidemic
World Health Organization (WHO) (and components)
Viral Hemorrhagic Fever
E. Coli

Infrastructure Security

Infrastructure security
Airport
CIKR (Critical Infrastructure & Key Resources)
AMTRAK
Collapse
Computer infrastructure
Communications infrastructure
Telecommunications
Critical infrastructure
National infrastructure
Metro
WMATA
Airplane (and derivatives)
Chemical fire
Subway
BART
MARTA
Port Authority
NBIC (National Biosurveillance Integration Center)
Transportation security
Grid
Power
Smart
Body scanner
Electric
Failure or outage
Black out
Brown out
Port
Dock
Bridge
Cancelled
Delays
Service disruption
Power lines

Southwest Border Violence

Drug cartel
Violence
Gang
Drug
Narcotics
Cocaine
Marijuana
Heroin
Border
Mexico
Cartel
Southwest
Juarez
Sinaloa
Tijuana
Torreon
Yuma
Tucson
Decapitated
U.S. Consulate
Consular
El Paso
Fort Hancock
San Diego
Ciudad Juarez
Nogales
Sonora
Colombia
Mara salvatrucha
MS13 or MS-13
Drug war
Mexican army
Methamphetamine
Cartel de Golfo
Gulf Cartel
La Familia
Reynosa
Nuevo Leon
Narcos
Narco banners (Spanish equivalents)
Los Zetas
Shootout
Execution
Gunfight
Trafficking
Kidnap
Calderon
Reyosa
Bust
Tamaulipas
Meth Lab
Drug trade
Illegal immigrants
Smuggling (smugglers)
Matamoros
Michoacana
Guzman
Arellano-Felix
Beltran-Leyva
Barrio Azteca
Artistic Assassins
Mexicles
New Federation

Terrorism

Terrorism
Al Qaeda (all spellings)
Terror
Attack
Iraq
Afghanistan
Iran
Pakistan
Agro
Environmental terrorist
Eco terrorism
Conventional weapon
Target
Weapons grade
Dirty bomb
Enriched
Nuclear
Chemical weapon
Biological weapon
Ammonium nitrate
Improvised explosive device
IED (Improvised Explosive Device)
Abu Sayyaf
Hamas
FARC (Armed Revolutionary Forces Colombia)
IRA (Irish Republican Army)
ETA (Euskadi ta Askatasuna)
Basque Separatists
Hezbollah
Tamil Tigers
PLF (Palestine Liberation Front)
PLO (Palestine Liberation Organization
Car bomb
Jihad
Taliban
Weapons cache
Suicide bomber
Suicide attack
Suspicious substance
AQAP (AL Qaeda Arabian Peninsula)
AQIM (Al Qaeda in the Islamic Maghreb)
TTP (Tehrik-i-Taliban Pakistan)
Yemen
Pirates
Extremism
Somalia
Nigeria
Radicals
Al-Shabaab
Home grown
Plot
Nationalist
Recruitment
Fundamentalism
Islamist

Weather/Disaster/Emergency

Emergency
Hurricane
Tornado
Twister
Tsunami
Earthquake
Tremor
Flood
Storm
Crest
Temblor
Extreme weather
Forest fire
Brush fire
Ice
Stranded/Stuck
Help
Hail
Wildfire
Tsunami Warning Center
Magnitude
Avalanche
Typhoon
Shelter-in-place
Disaster
Snow
Blizzard
Sleet
Mud slide or Mudslide
Erosion
Power outage
Brown out
Warning
Watch
Lightening
Aid
Relief
Closure
Interstate
Burst
Emergency Broadcast System

Cyber Security

Cyber security
Botnet
DDOS (dedicated denial of service)
Denial of service
Malware
Virus
Trojan
Keylogger
Cyber Command
2600
Spammer
Phishing
Rootkit
Phreaking
Cain and abel
Brute forcing
Mysql injection
Cyber attack
Cyber terror
Hacker
China
Conficker
Worm
Scammers
Social media

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.


#bostonstrong
explosion
marathon
brothers
monday
search
cooker
#marathonmonday
chechen
response
#bostonmarathon
strong
official
overcome
wheelchair
victim
boat
hurt
race
terrorist
city
blue
running
fund
child
safe
injured
scanner
#wewillneverforget
what
attack
together
shoes
usa
intense
manhunt
injury
strength
horrifying
deaths
news
#staystrong
athletes
historic
copley
investigation
where
prayers
okay
shootoutarrest
dzhokhar
police
community
swat
collier
guns
people
keep
restrict
dead
madness
fear
drink
amputee
fire
hiding
blast
love
runner
threat
fight
kill
harvard
pride
boylston
officer
explode
sad
drive
allston
culprit
sean
again
hero
emt
russian
mit
suspect
diversion
rescue
casualties
army
brave
rip
legs
april
rice
hospital
building
hunt
run
find
shocking
support
mbta
help
ambulance
zubeidat
radio
god
pray
qaeda
dartmouth
mgh
lawyer
somerville
stand
#staysafe
horror
feet
lockdown
memorial
detective
save
tragedy
scary
unexpected
family
debris
yellow
one
pressure
terrorism
weather
children
leaves
drunk
danger
runners
shutdown
shooting
tsarnaev
tamerlan
chase
line
watertown
survivors
backpack
finish

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)

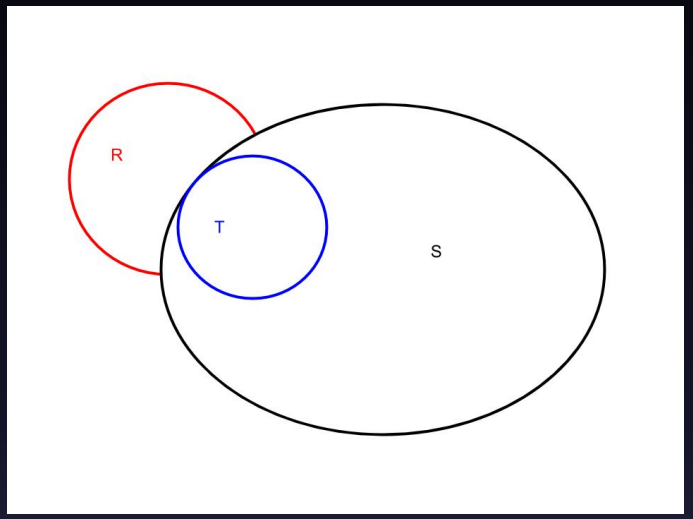
Thresher keeps humans in the loop and helps them find more and better words faster.

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.

Target set words

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,…

Non-target set words

game, red sox, bruins, celtics, back, tonight, fan, #mlb, night, chicago, new york, garnett, fenway, rondo, #job, playoff, yankees, blackhawks, stanley, pizza, #nhl, draft,…

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: الاستشهادية عمليات **(martyrdom operations)**
from "Haqibatu'l-Mujahid" (A Mujahid's Bookbag)

S: "Pulpit of Tawhid and Jihad"
A Jihadist web library

العدو	**enemy**
قتل	**killing**
والنكاية	**to vex or spite ("vex the infidels")**
يَعْلَمُهُمْ	**teach them**
الْخَيْلِ	**steed**
وَأَعِدُّوا	**fight**
تُظْلَمُونَ	**wronged**
ترهبون	**terrify**
الغلام	**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?

Egyptian	Gulf	Levantine	MSA	Translation
ايه	وش	شو	ماذا	**What**
معرفش	مادري	ما بعرف	لا يعرف	**Don't know**
عايز	يبي	بدي		**To want/I want**
عاوز	يبغى	بدك		**To want/You want**
عايزين	يبون	بدكم		**To want/You want (pl.)**

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