Text as Data

Juraj Medzihorsky

CEU CENTRAL EUROPEAN UNIVERSITY

2017-03-14

Programming

The Talent Myth

Reality


DRAPER
EXPERT

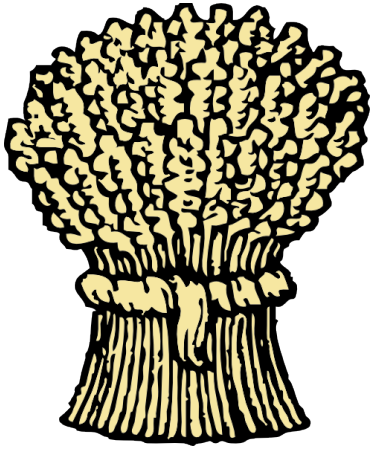
Document Scraping

Document Scraping

- Numbers and text in files
- Local
- Web

Web

- eXtensible Markup Language (XML)
- APIs (e.g. for Twitter)



Text Analysis

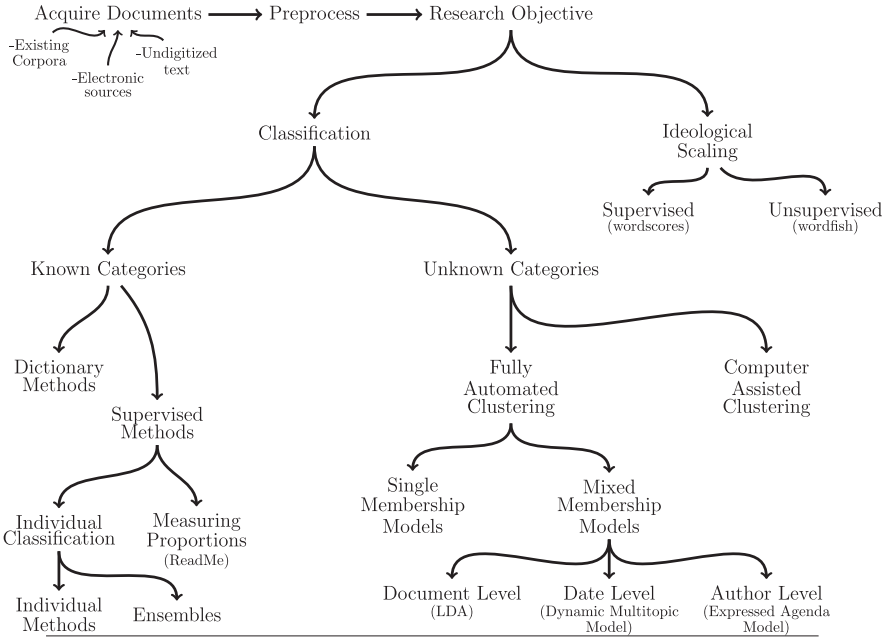


Fig. 1 An overview of text as data methods.

Dictionary-Based Methods

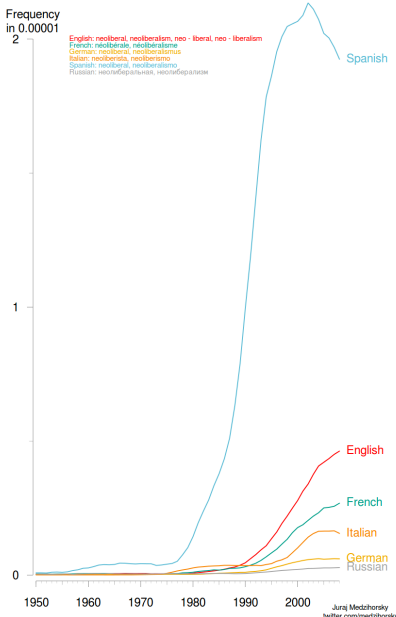
Google N-Grams

Dictionary-Based Methods

- Build a dictionary
- Statistical models are not necessary

Figure 1: Frequency of appearance of “Expertise” and “Expert” in Google Books from 1800 to 2000

'Neoliberalism' in Google Book Corpora



Script: ngramr

Scaling

Scaling Goals

- One or more dimensions
- Place documents (texts, speeches) in a space
- Place words in the same space

Common Scaling Methods

Supervised: Dictionary-Based*

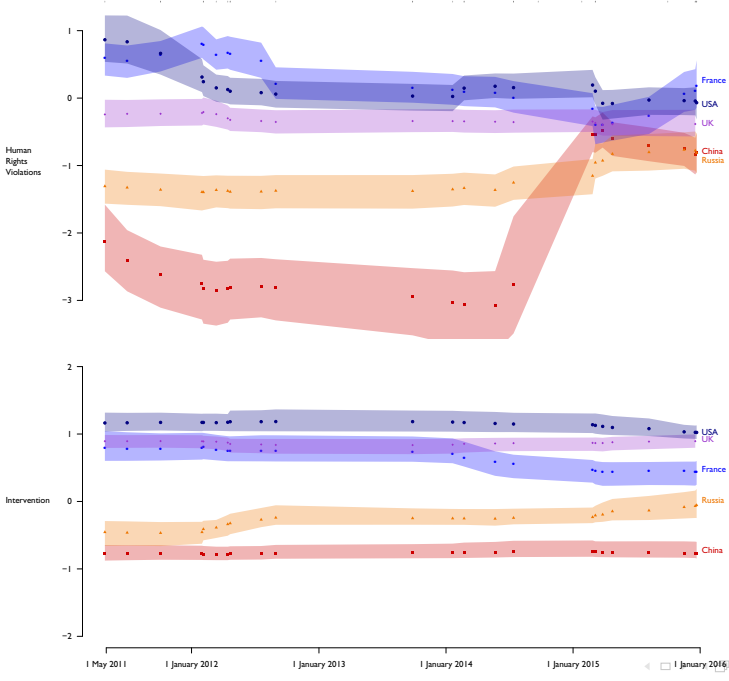
Supervised: Wordscores

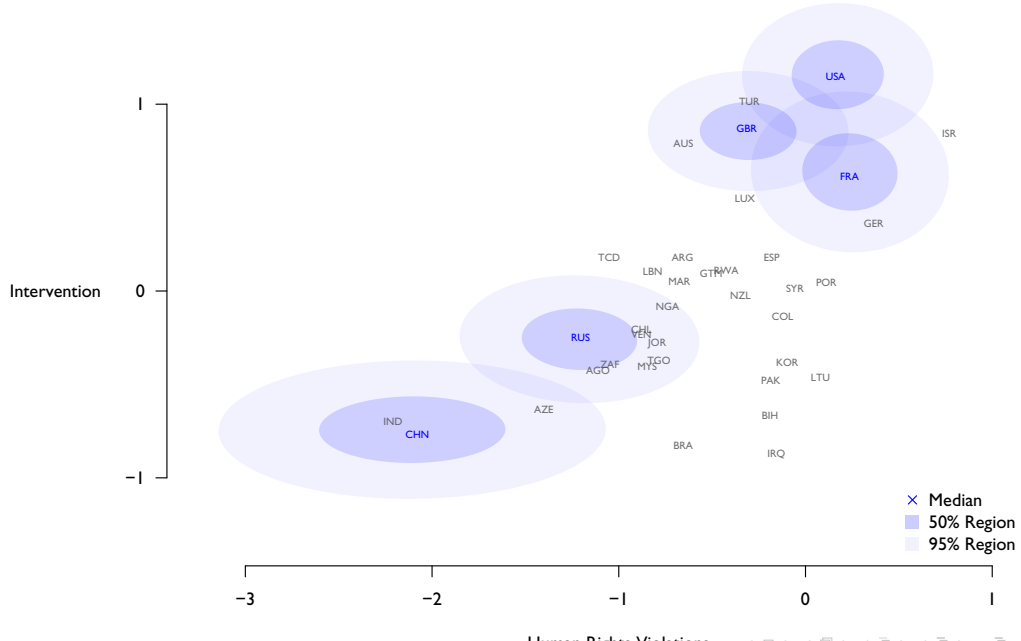
Unsupervised: Wordfish

Unsupervised: Correspondence Analysis

Dictionary-Based Scaling: UNSC Speeches on Syria

Theme	Words
No HR Violations	conflict, violence, tension, struggle, war, stability, destabilize, security, crisis, escalate, incite, threat, chaos, cycle, fighting, casualties, losses, parties, clash, dispute
HR Violations	repression, humanity, crime, moral, torture, persecution, abuse, oppress, repress, life, incite, tyranny, terrorism, children, women, perpetrator, victims, accountable, massacre, crackdown, targeting, indiscriminate, brutal, barrel, genocide, cleansing, school, hospital, kill
Anti-intervention	process, charter, implementation, dialogue, constructive, consensus, diplomatic, reconciliation, settlement, comprehensive, inclusive, mediation, effort, negotiation, proposal, solution
Pro-intervention	urgent, action, assistance, support, aid, sanctions, arrest, stop, intervention, end, deliver


Human Rights Violations
1
0
−1
−2
−3
France
USA
UK
China
Russia
Intervention
2
1
0
−1
−2
USA
UK
France
Russia
China
1 May 2011
1 January 2012
1 January 2013
1 January 2014
1 January 2015
1 January 2016


USA
TUR
GBR
ISR
AUS
FRA
LUX
GER
TCD
ARG
ESP
LBN
RWA
GTM
MAR
SYR
POR
NZL
NGA
COL
CHL
RUS
YEM
JOR
TGO
ZAF
AGO
MYS
KOR
PAK
LTU
AZE
BIH
IND
CHN
BRA
IRQ
Intervention
1
0
−1
Median
50% Region
95% Region
−3
−2
−1
0
1
Human Rights Violations

Script: UNSC

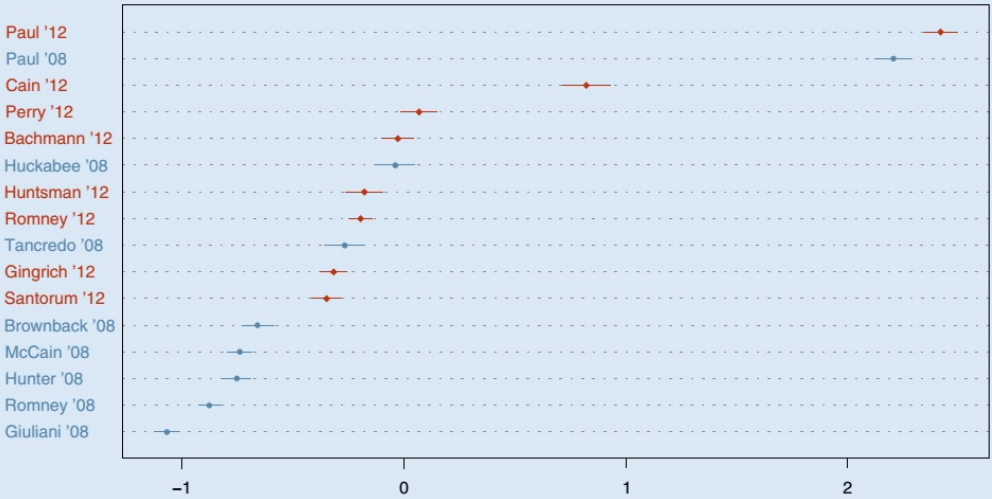
Unsupervised Scaling: US Presidential Primaries

US Presidential Primaries

- 2008 and 2012 GOP presidential primaries
- Debate transcripts from a UCSB website
- Expected move towards Tea Party positions
- Unsupervised scaling: Wordfish

Figure 1

Candidate Positions



Candidate positions extracted from their pre-Iowa debate speeches with bootstrapped 95% confidence intervals (1,000 replications). 2008 candidates denoted by circles (blue) and 2012 candidates by diamonds (red).

Table 4

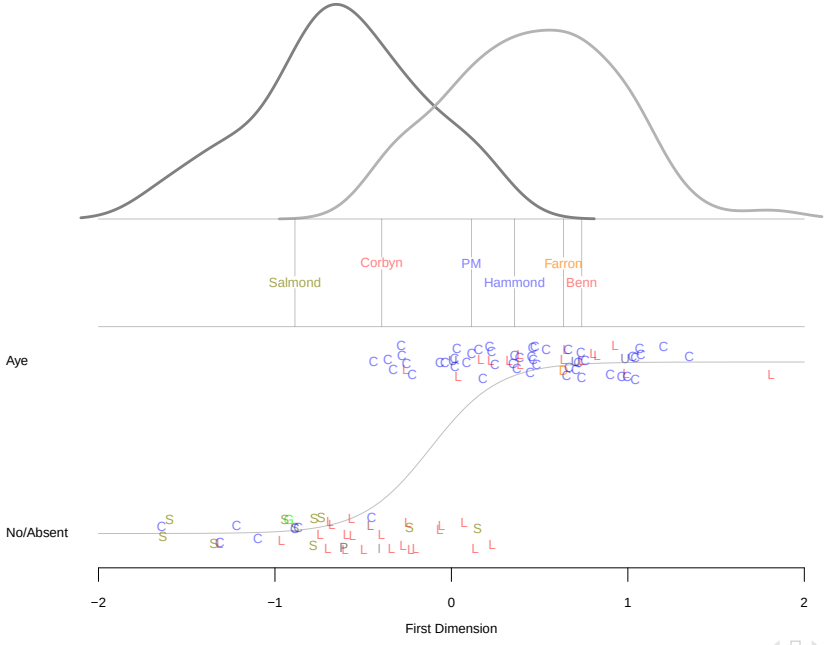
Selected Five-Sentence Sequences Spoken by a Single Candidate in a Single Debate

NEGATIVE, −1±0.1	POSITIVE, +1±0.1
"I have joined together across the aisle on a number of pieces of legislation, many of them very important. I'm proud of my legislative record of conserving my ideals and my conservative principles and getting things done in Washington. And I am proud of that, and I will continue to hold to those ideals. But I will reach across the aisle to the Democrats who I have worked with, who know me, and we know we can work together for the good of this country. Let's raise the level of dialogue and discussion and debate in this campaign." (McCain on December 12, 2007; score: −1.1)	"Repeal Dodd-Frank, repeal Obamacare. It really isn't that tough if you try. It is easy to turn around this economy, just have the backbone to do it. Well, as president of the United States, I would not be reappointing Ben Bernanke, but I want to say this. During the bailout, the $700 billion bailout, I worked behind the scenes against the bailout, because one of the things that I saw from the Federal Reserve, the enabling act legislation is written so broadly that, quite literally, Congress has given the Federal Reserve almost unlimited power over the economy." (Bachmann on September 12, 2011; score: 1.9)
"It's the one place I found to agree with President Obama. If every parent in America had a choice of the school their child went to, if that school had to report its scores, if there was a real opportunity, you'd have a dramatic improvement. I visited schools where, three years earlier, there were fights, there were dropouts, there was no hope. They were taken over by a charter school in downtown Philadelphia, and all of a sudden the kids didn't fight anymore, because they were disciplined. They were all asked every day, what college are you going to? Not are you going to go to college, what college are you going." (Gingrich on September 7, 2011; score: −1)	"If we look for it, you'll realize that our national sovereignty is under threat. Yes, and I would like to state that, to the statement earlier made that we all went to Washington to change Washington and Washington changed us, I don't think that applies to me; Washington did not change me. I would like to change Washington, and we could by cutting three programs, such as the Department of Education–Ronald Reagan used to talk about that–Department of Energy, Department of Homeland Security is the biggest bureaucracy we ever had. And besides, what we can do is we can have a stronger national defense by changing our foreign policy. Our foreign policy is costing us a trillion dollars, and we can spend most of that or a lot of that money home if we would bring our troops home." (Paul on November 28, 2007; score: 2)
"I can tell you a good union, the Steel Workers Union. When last year, Chris, we had a strike in a Kansas plant that made the tires for our humvees, I called up the president of the Steelworkers and the president of Goodyear, and within a very short period of time, they were working together, they got that thing done for the good of the country. A union is a receptacle of power, just like management. But those folks love this country, they love their family, and they helped to build a middle class, which has been important for America and for our party. We need to work with unions to win this presidency." (Hunter on October 9, 2007; score: −0.9)	"There's a responsible way for the federal government to do the things that it should do. Running organizations like the TSA, I would agree with Representative Paul, no. Having the federal government responsible for trying to micromanage Medicare, no, trying to micromanage education, no. The federal government is not good at micromanaging anything. This is why I believe in empowering the states to do more and limit what the federal government does with regard to those kinds of program." (Cain on August 11, 2011; score: 2.1)

Sequences were selected from all such sequences longer than 500 characters and within ±0.1 to two points on the dimension: −1 and +2. Sequences in the first column characterize the negative end of the extracted dimension and sequences in the second column the positive end.

Script: Primaries

Unsupervised Scaling: UK Parliament Speeches on Syria Airstrikes

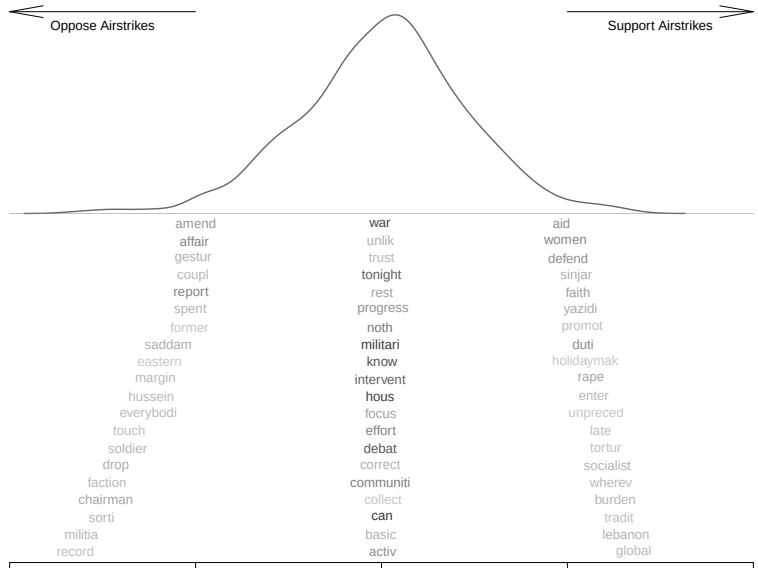


British airstrikes in Syria were approved by the Parliament on 2 December 2015. Over 100 MPs spoke in the debate. This figure shows the first dimension extracted from their speeches with unsupervised scaling. The top panel shows gaussian kernel density estimates of scores by support for the airstrikes. The bottom panel shows the fit of a generalized linear model with Bernoulli sampling distribution and logistic link function. In the model, the score is the sole predictor of the support for the airstrikes.

Party

- C Conservative
- D Liberal Democrats
- G Green
- I Independent
- L Labour/Co−operative
- L Labour
- L Social Democratic and Labour
- P Plaid Cymru
- S Scottish National
- U Democratic Unionist
- U Ulster Unionist

Juraj Medzihorsky
twitter.com/medzihorsky



British airstrikes in Syria were approved by the Parliament on 2 December 2015. Over 100 MPs spoke in the debate. This figure shows the first dimension extracted from their speeches with unsupervised scaling. Higher scores on the dimension are associated with voting for the airstrikes. The upper panel shows a gaussian kernel density estimate of the stem scores. The bottom panel shows the scores of sixty selected stems. The stems on the left and right have the lowest and highest scores, respectively. In the middle are the stems least associated with the dimension.

Stem Frequency

a 1000
a 500
a 100
a 50
a 10
a 5

Juraj Medzihorsky
twitter.com/medzihorsky

‘Topic’ Models

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

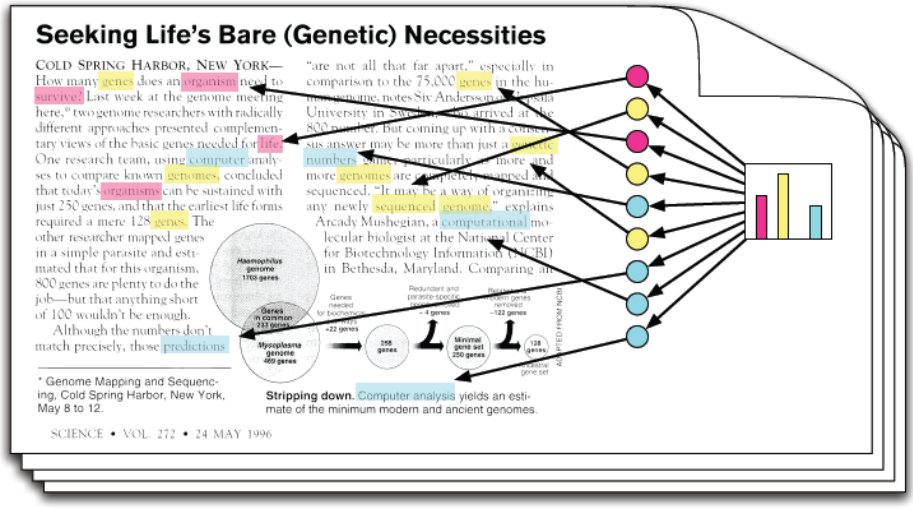
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



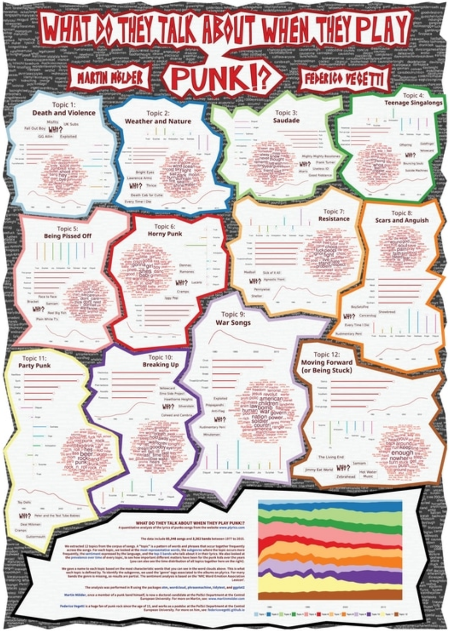
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an-

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.


WHAT DO THEY TALK ABOUT WHEN THEY PLAY PUNK!?
MARTIN MÖLDER
FEDERICO VEGETTI
Topic 1: Death and Violence
Topic 2: Weather and Nature
Topic 3: Saudade
Topic 4: Teenage Singalongs
Topic 5: Being Pissed Off
Topic 6: Horny Punk
Topic 7: Resistance
Topic 8: Scars and Anguish
Topic 9: War Songs
Topic 10: Breaking Up
Topic 11: Party Punk
Topic 12: Moving Forward (or Being Stuck)

Script: STM