Information Storage and Retrieval

CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

Text Analytics: Topic Models
11 April 2017

Text Analytics

Data Mining View: Explore patterns in textual data

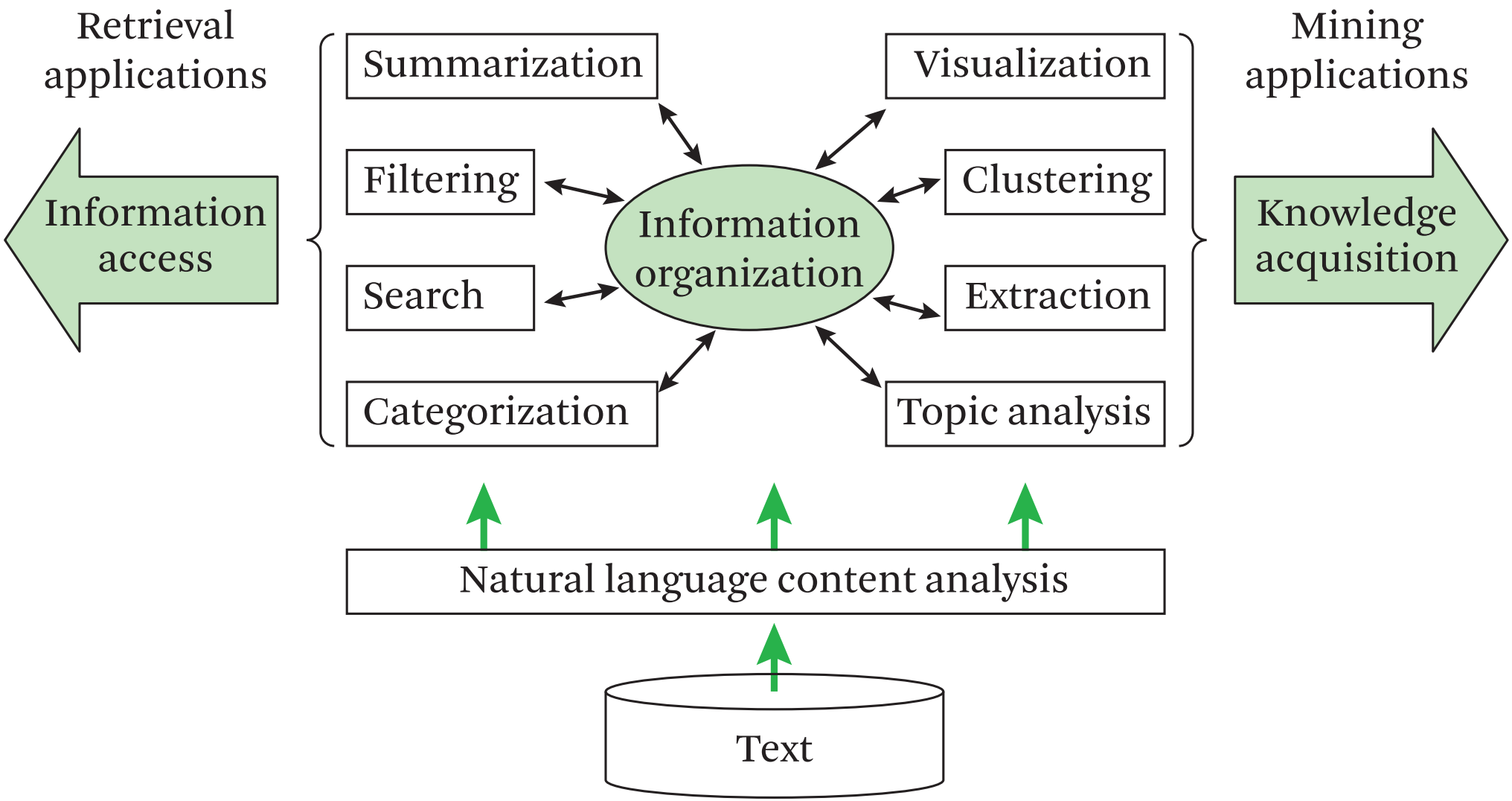
- Find latent topics
- Find topical trends
- Find outliers and other hidden patterns

Natural Language Processing View: Make inferences based on partial understanding of natural language text

- Information extraction
- Question answering

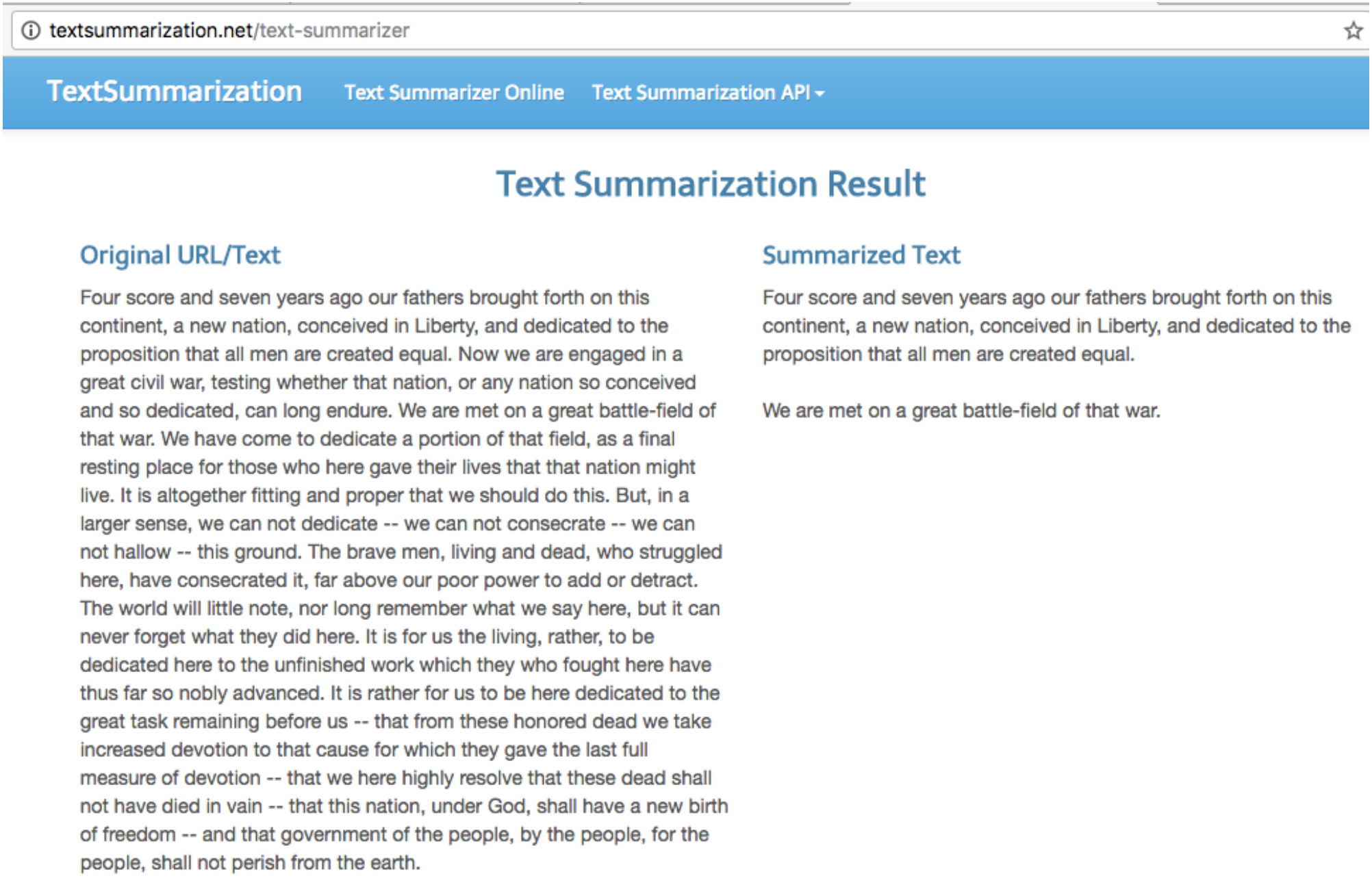
Zhao and Massung. Text Data Management and Analysis (2016)

Text Analytics



Zhao and Massung. Text Data Management and Analysis (2016)

Summarization



Text Categorization

Google Corporation®
123 Buckingham Palace Road
London SW1W 9SH
United Kingdom

Good day Sir/Madam.

You have successfully been picked as one of our 12 Lucky Winners in this months Lottery Draw, Please see attached file for more details.

Best regards
Sundar Pichai
CEO Google Inc.
sundarpichai@gpromo-team.com

Text Categorization

The 2nd edition of iOS Apps for Masterminds is now available. This is the first book to teach how to program applications for iOS 10 with Swift 3 and Xcode 8.

iOS Apps for Masterminds is great for a semester course, a complement to an undergraduate program, or to get you familiar with the technology. The book gradually introduces the concepts to guide students step by step on how to create an application from scratch. It goes from basic programming concepts like variables and memory management to more complex subjects like Core Data and iCloud (more than 800 pages of content).

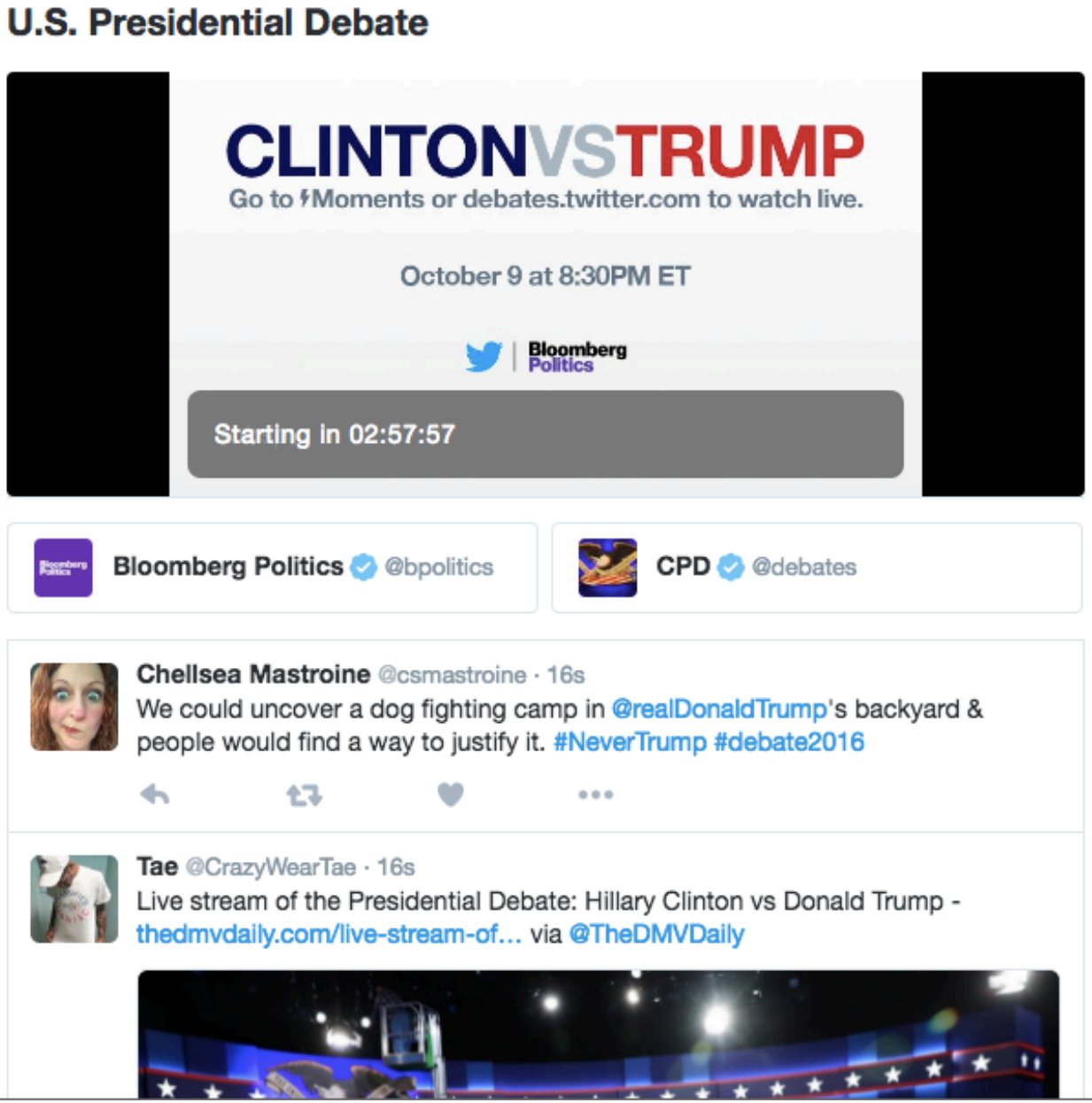
Read the preview at Amazon
www.amazon.com/dp/1537517880/

The For Masterminds books are also the less expensive and more comprehensive books on the subjects. Perfect for a low-budget course.

To see the Table of Contents and examples included in the book, visit our website
www.formasterminds.com

For additional information or promo codes to download the eBook for free, please reply this email or contact me at info@jdgauchat.com.

Text Categorization



Text Categorization

How would you represent each "document"?

How would you compare them?

What approach ("classifier") would you use?

How would you evaluate your approach?

Text Categorization

Other examples?

Text Categorization

Other examples?

Authorship attribution

Positive or negative movie review

Age / gender identification

…

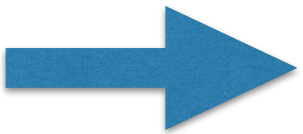
Document Clustering

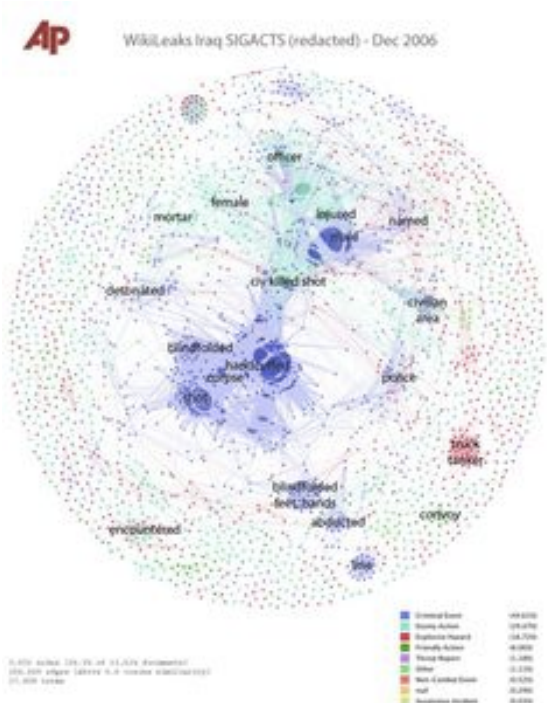
Suppose you're an analyst for:

GoldmanSachs looking for **major shifts in a company** (e.g., Theranos)

the **NSA** looking for **evidence of a terror cell** operating in a particular country

… ?





Document Clustering

How would you represent each "document"?

How would you compare them?

What approach ("clustering algorithm") would you use?

How would you evaluate your approach?

Information Extraction

Find and understand limited relevant parts of texts

Gather information from many pieces of text

Produce a structured representation of relevant information:

- relations (in the database sense), a.k.a., a knowledge base

Goals:

- Organize information so that it is useful to people
- Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Sherry Escalante <s-escalante@tamu.e Oct 7 (2 days ago)

to AM-ENGINEER-FA.

Engineering Faculty,
As you know, service on a faculty search committee requires you to take the training session offered by the Dean of Faculties. They will be offering a session for engineering faculty and staff on **Wednesday, October 26th from 2:00-4:00 pm** in room 217 of the Civil Engineering Office Building (CEOB). Add to Calendar, please register here by 10/21 :

Google

what year was texas founded

All News Images Shopping Maps More Search tools

About 7,490,000 results (0.69 seconds)

Texas / Founded

December 29, 1845

Feedback

People also ask

What is the history of Texas?

When was Texas first settled?

What is the state of Texas?

Why is the capital of Texas Austin?

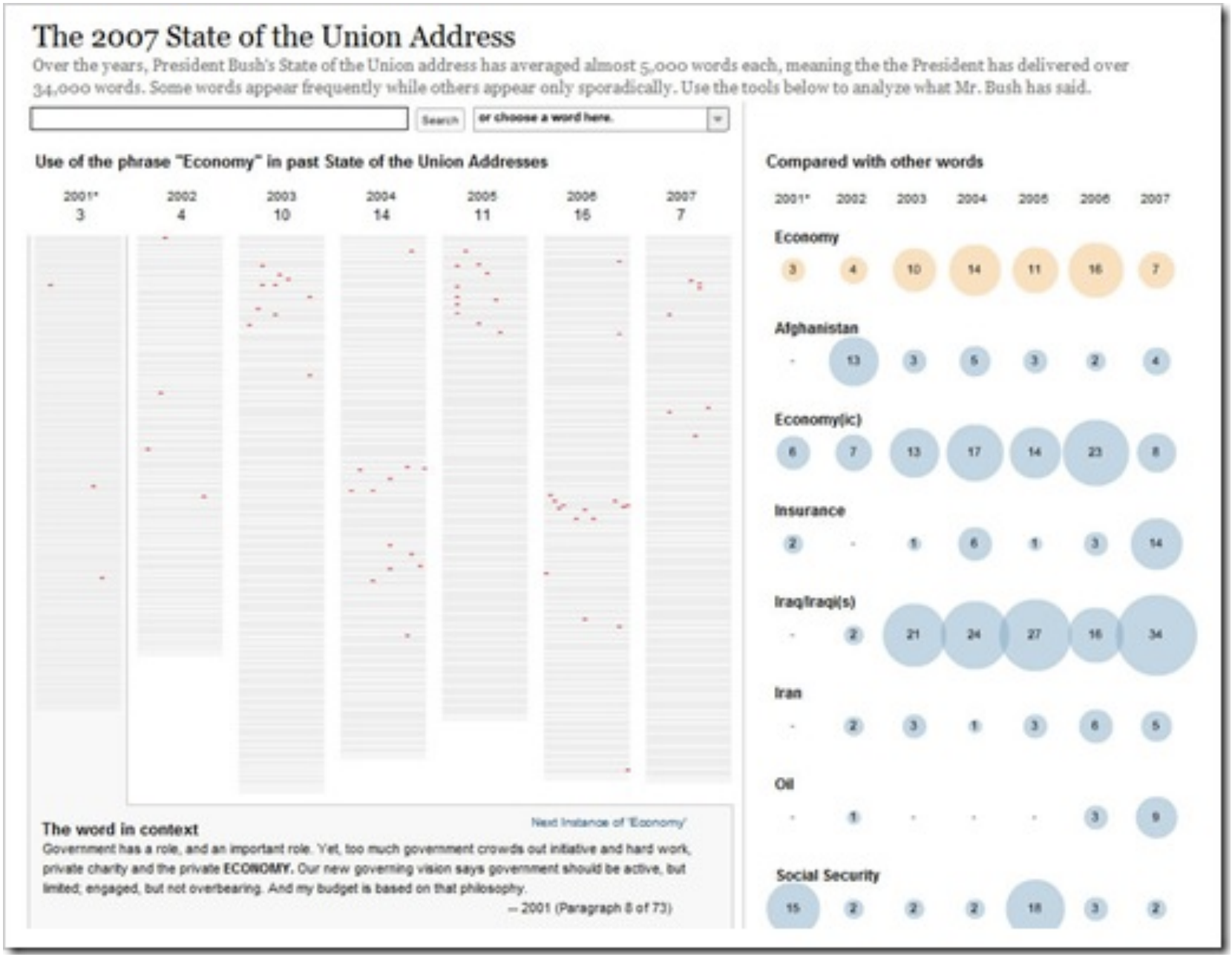
Texas - Wikipedia, the free encyclopedia

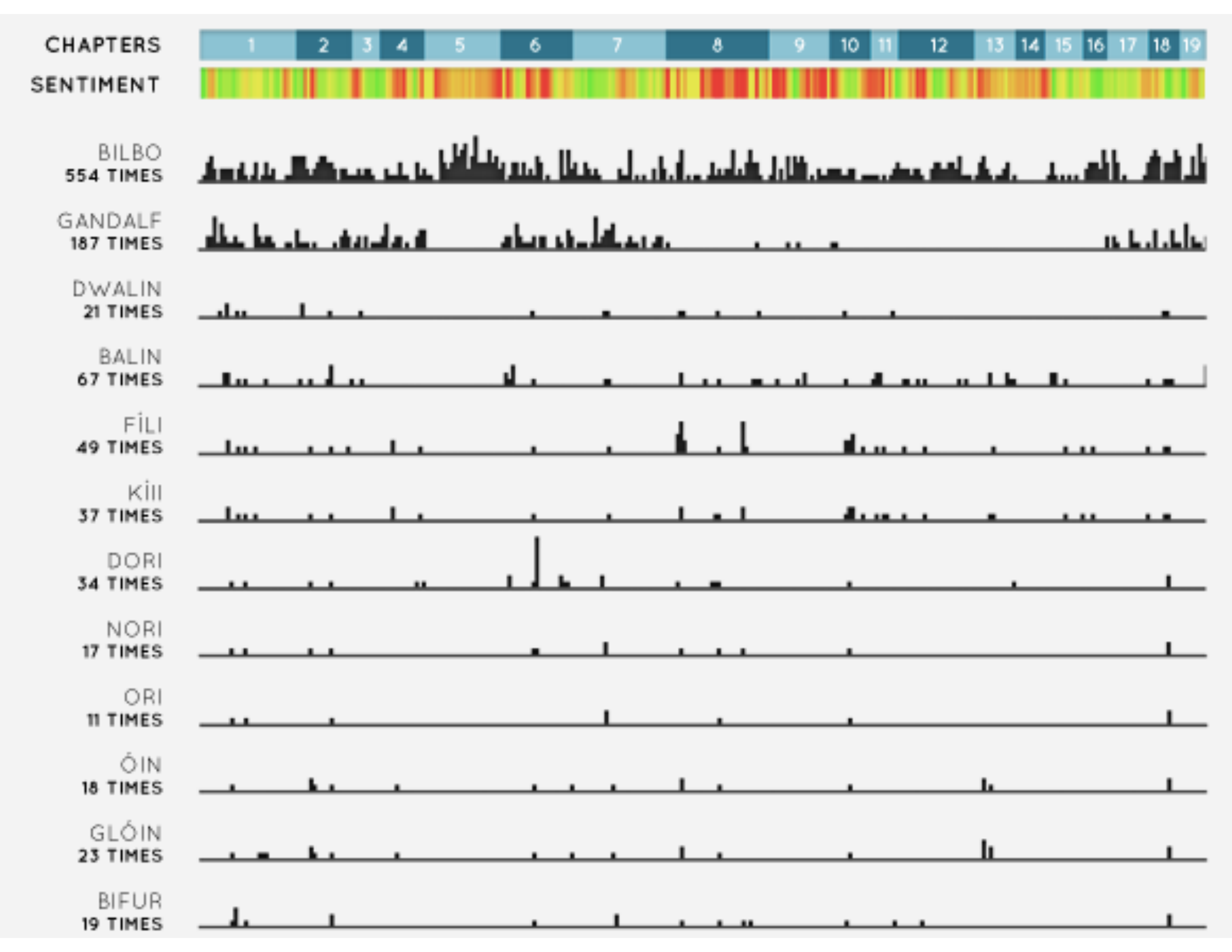
https://en.wikipedia.org/wiki/**Texas** Wikipedia

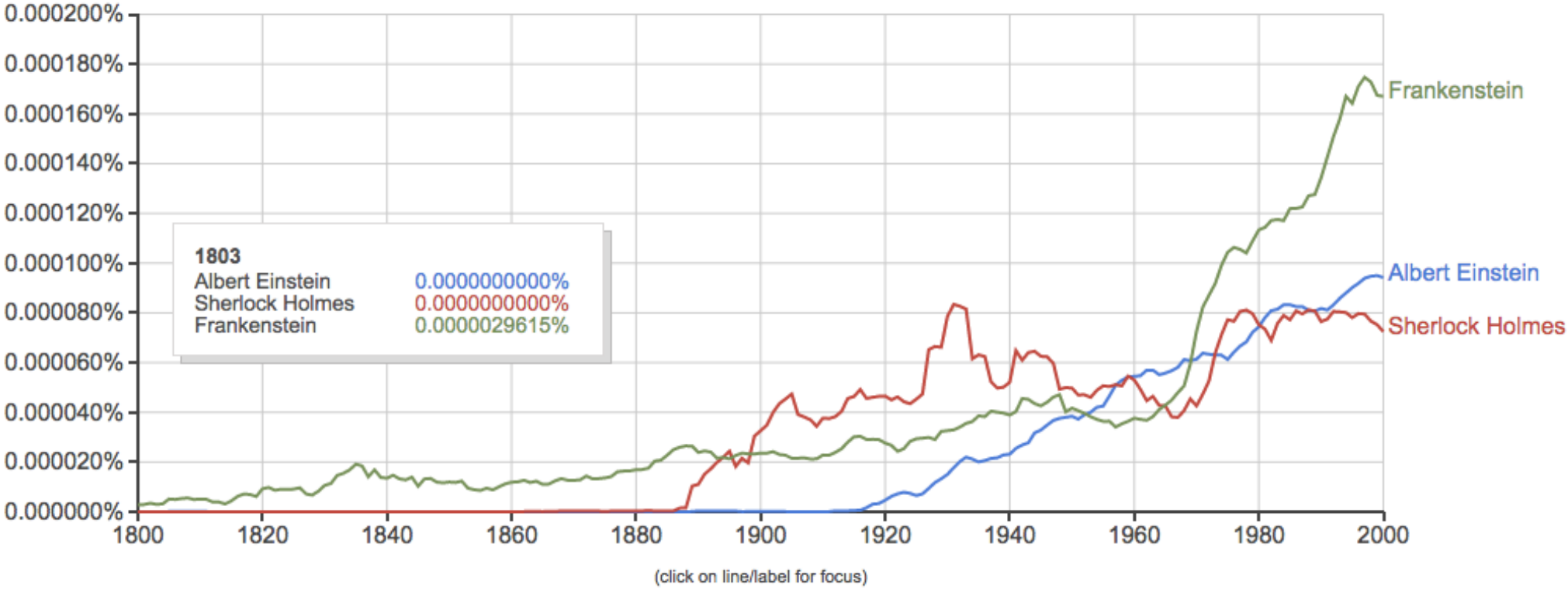
Information Extraction

How would you approach this problem?

Text Visualization




CHAPTERS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
SENTIMENT
BILBO
554 TIMES
GANDALF
187 TIMES
DWALIN
21 TIMES
BALIN
67 TIMES
FİLI
49 TIMES
KİII
37 TIMES
DORI
34 TIMES
NORI
17 TIMES
ORI
11 TIMES
ÔIN
18 TIMES
GLÓIN
23 TIMES
BIFUR
19 TIMES



https://books.google.com/ngrams

English

Version 20120701
total_counts

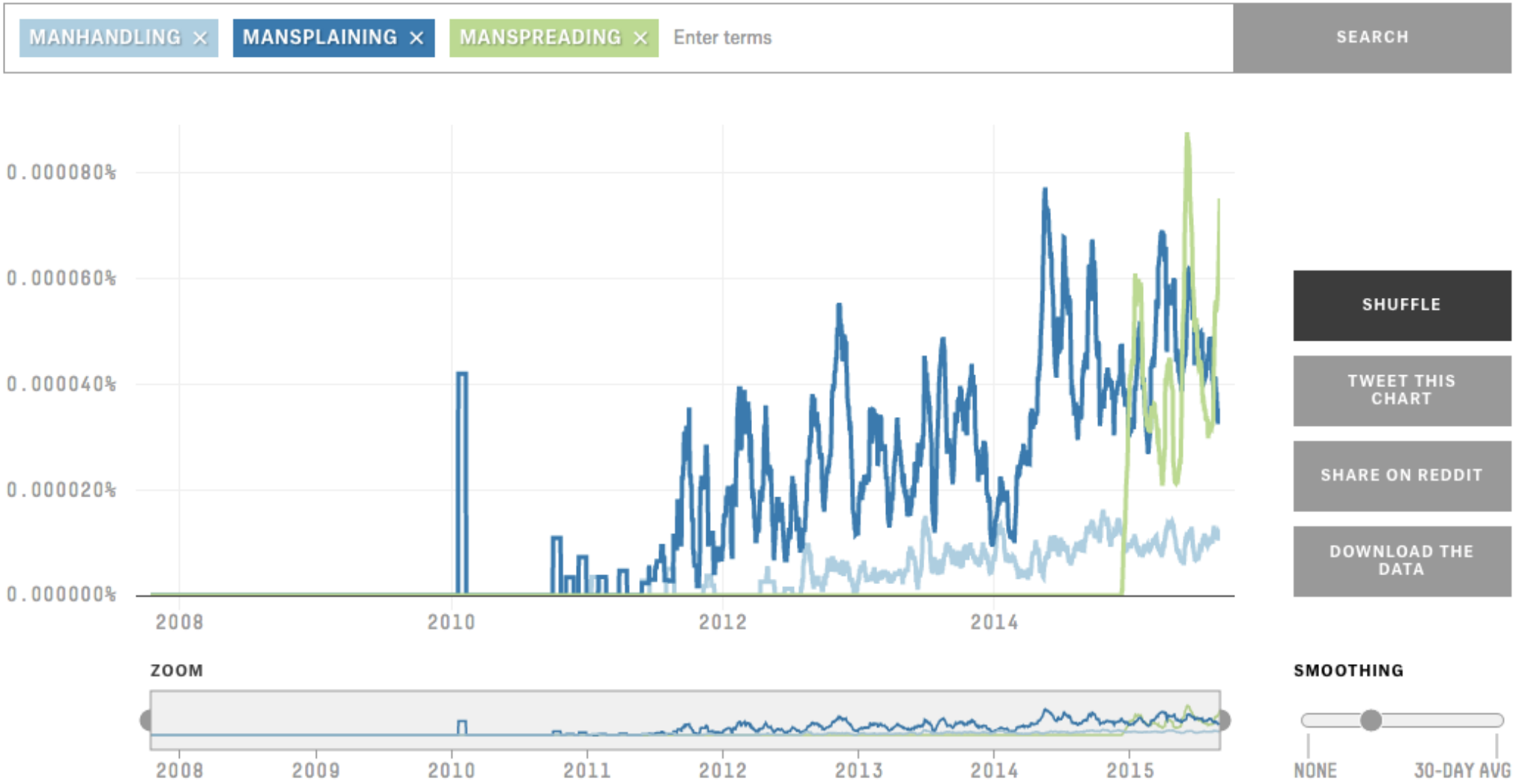
1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

2-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

3-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

4-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb

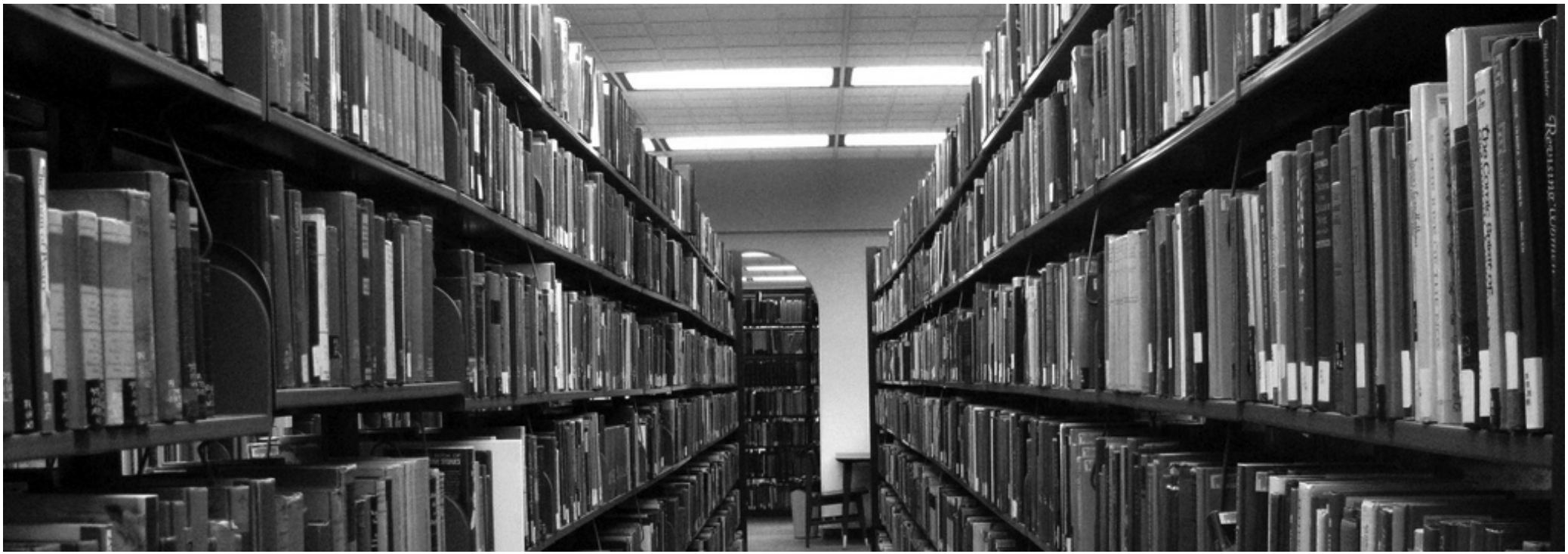
http://storage.googleapis.com/books/ngrams/books/datasetsv2.html



http://projects.fivethirtyeight.com/reddit-ngram/

Topic Models

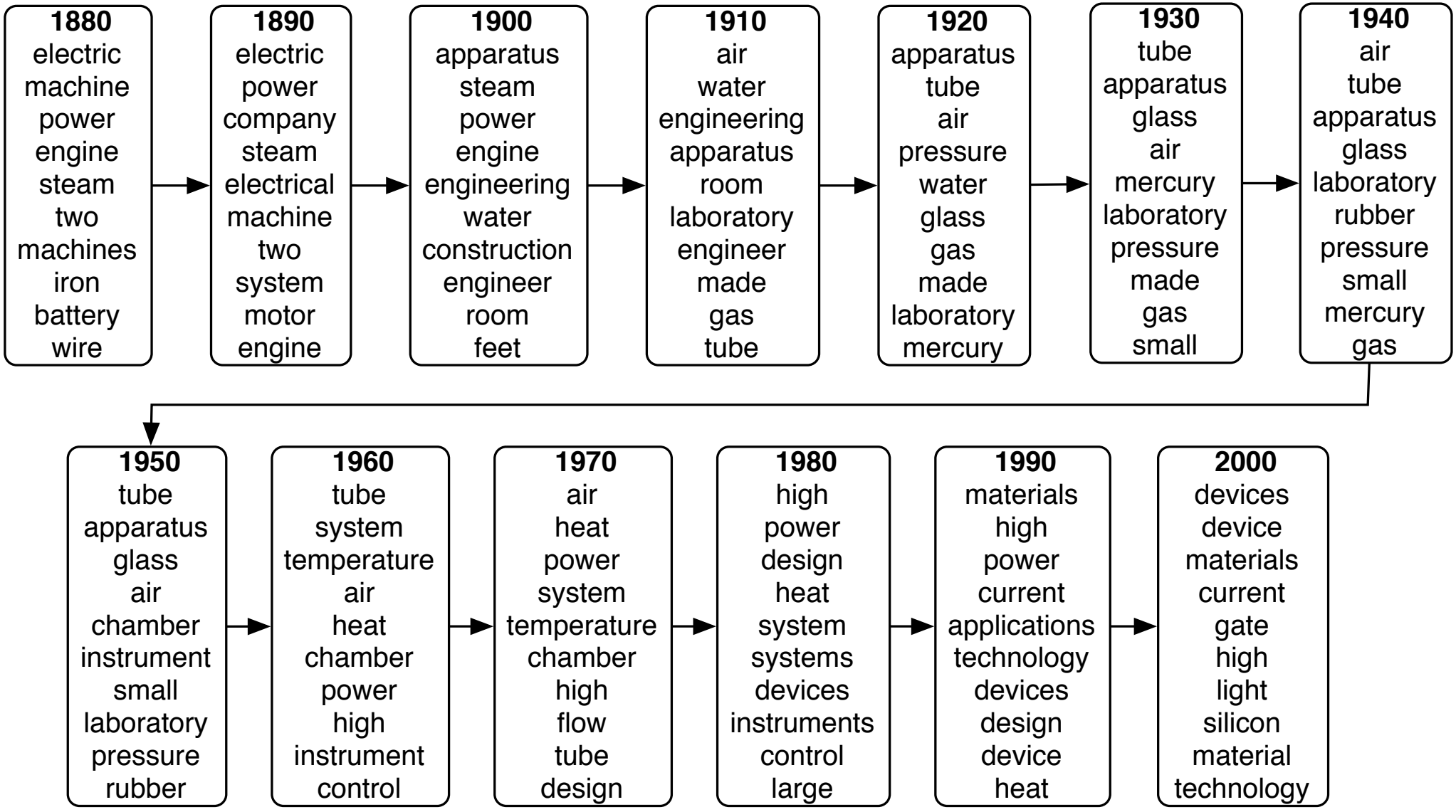
http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf



Input: An unorganized collection of documents

Output: An organized collection, and a description of how

Blei. Probabilistic Topic Models: Origins and Challenges (2013 Topic Modeling Workshop at NIPS)


1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire
1890
electric
power
company
steam
electrical
machine
two
system
motor
engine
1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet
1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube
1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury
1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small
1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas
1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber
1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design
1980
high
power
design
heat
system
systems
devices
instruments
control
large
1990
materials
high
power
current
applications
technology
devices
design
device
heat
2000
devices
device
materials
current
gate
high
light
silicon
material
technology

LDA = Latent Dirichlet Allocation

by Blei, Ng, and Jordan at Berkeley

introduced in 2003

hugely influential

tons of implementations

Dave Blei: Professor at Columbia, ACM Fellow

Andrew Ng: Professor at Stanford, founded Google Brain, founded Coursera, formerly Chief Scientist at Baidu

Michael Jordan: Professor at Berkeley, hugely influential, (dad is an Aggie!)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

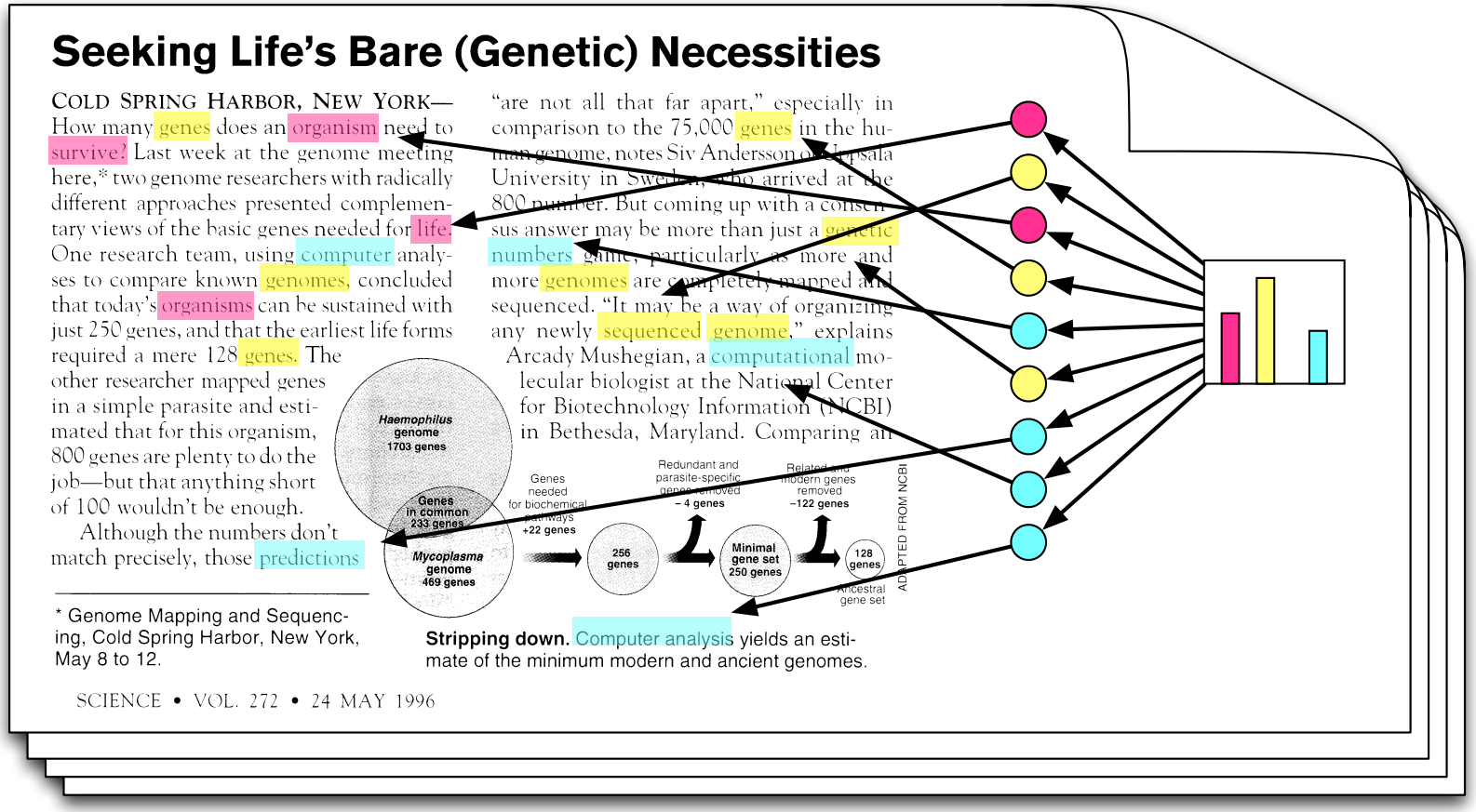
life	0.02
evolve	0.01
organism	0.01
...	

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

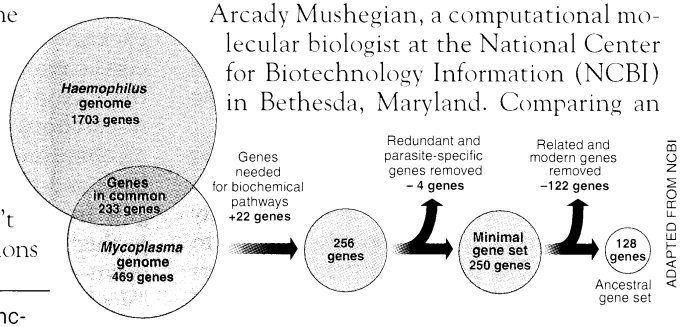
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

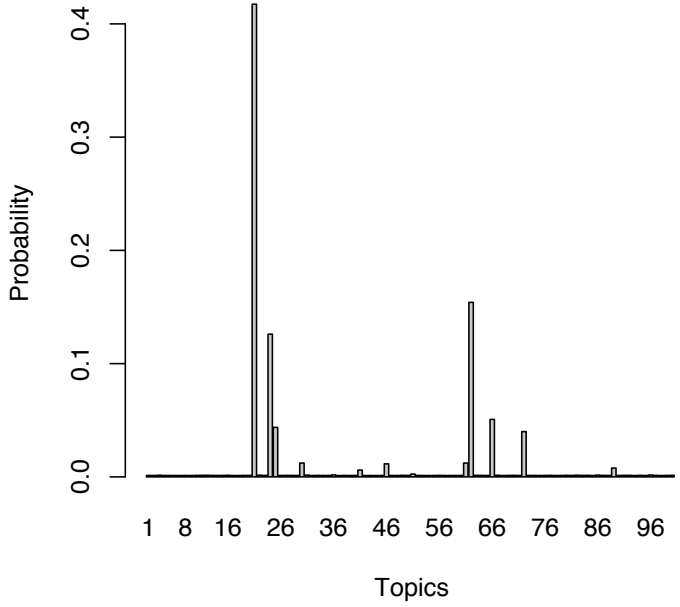
Seeking Life's Bare (Genetic) Necessities

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.

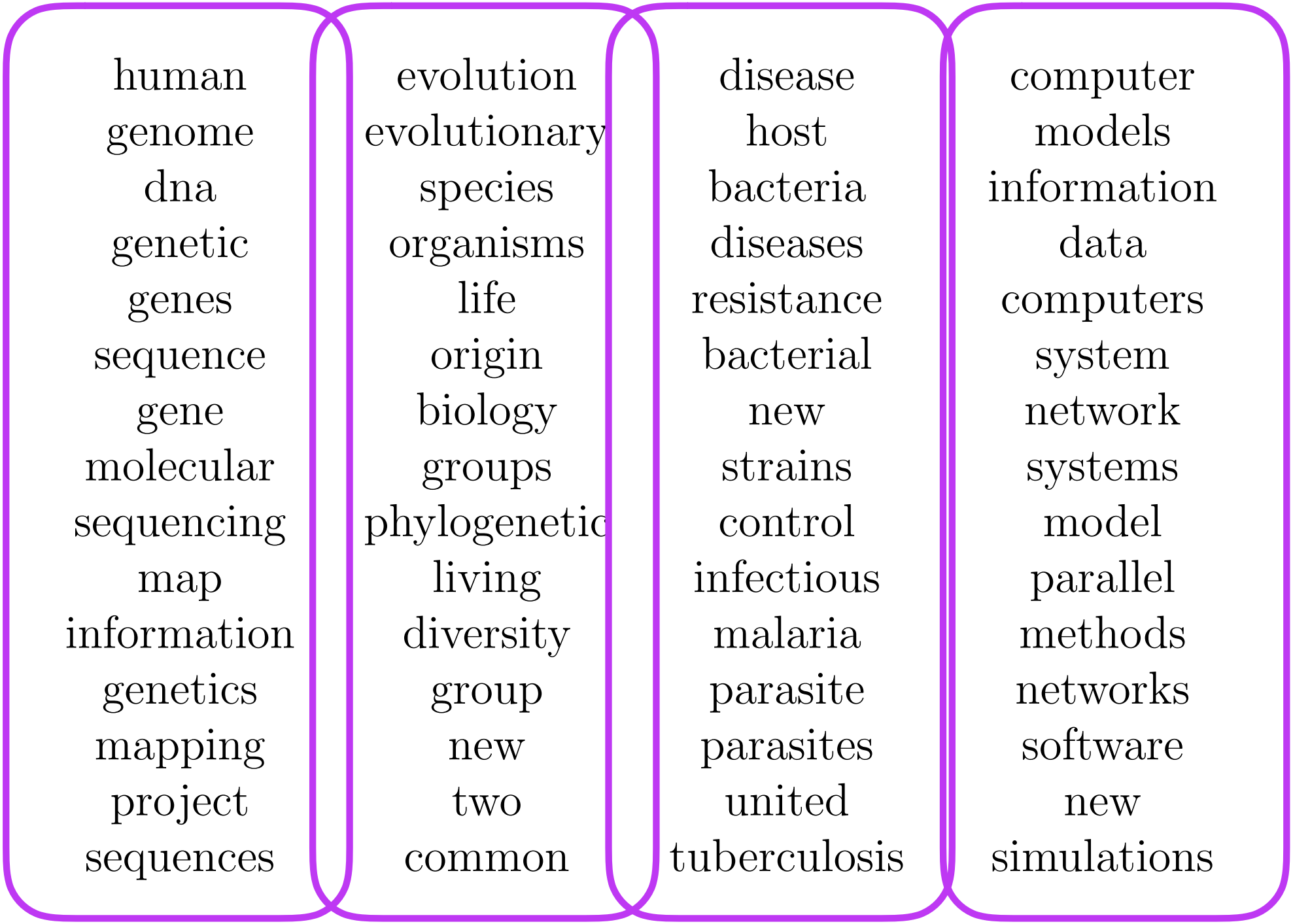
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



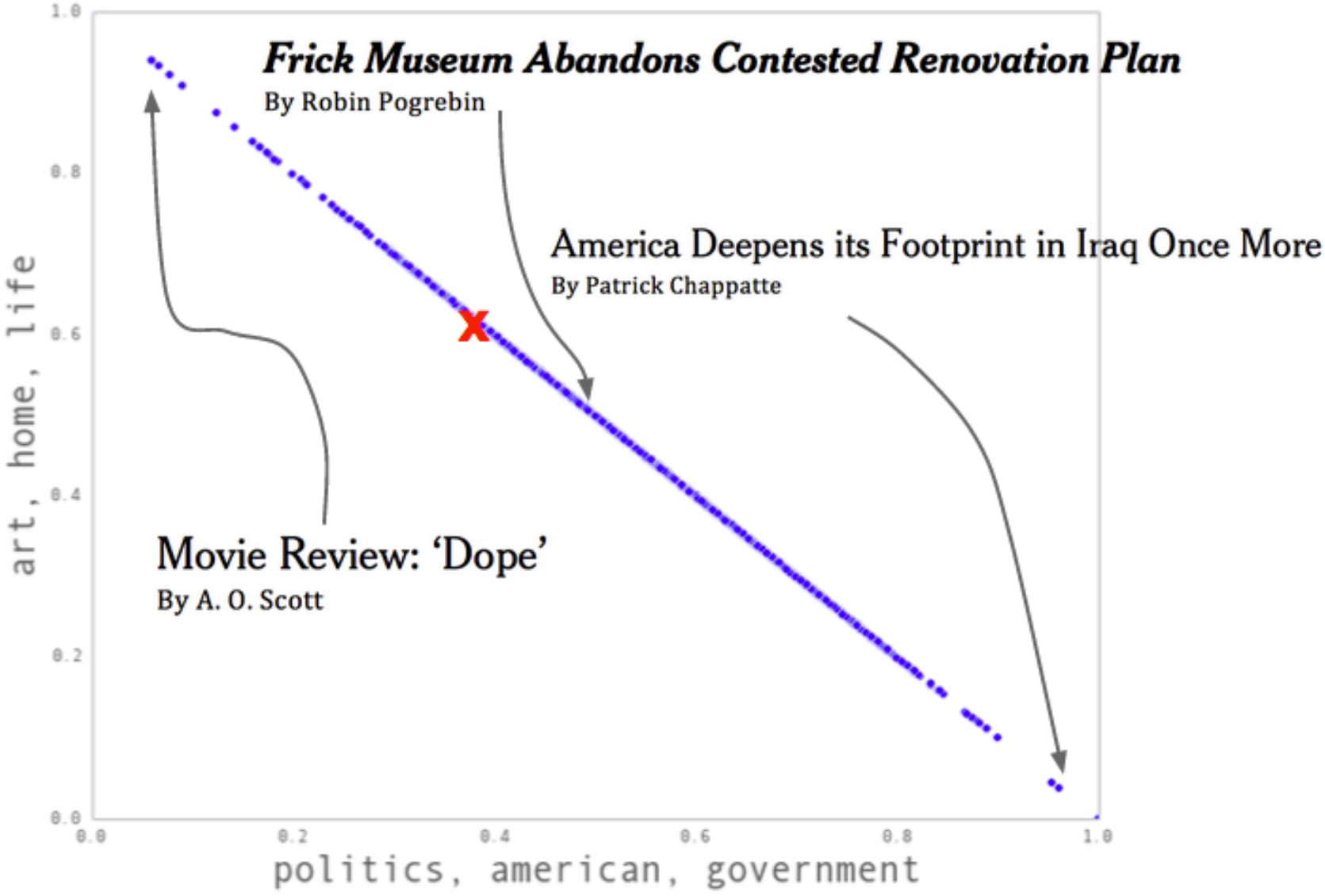
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.



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


human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences
evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common
disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis
computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

LDA in action …

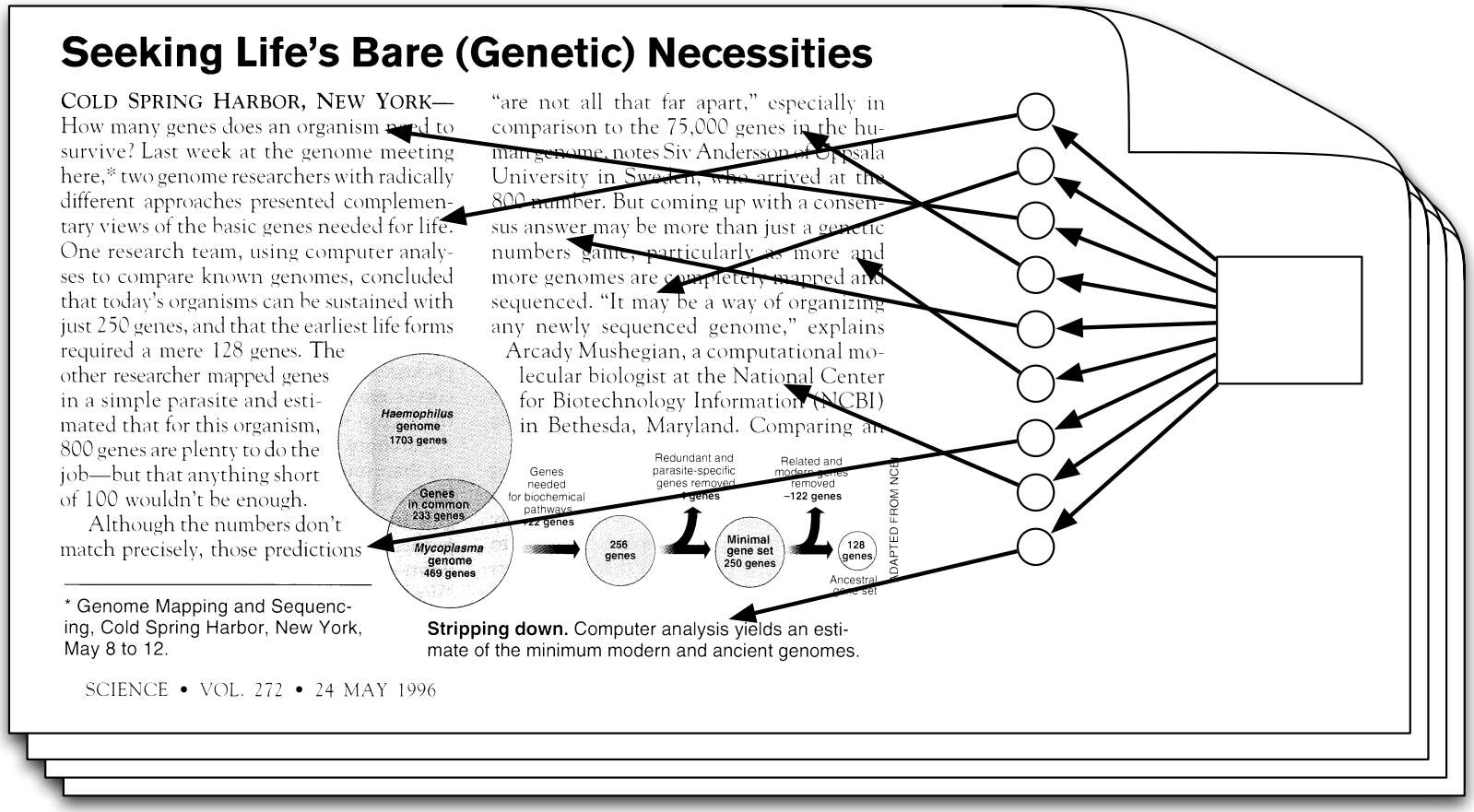


http://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/

Topics

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

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

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

Consider the following "documents"

I like to eat broccoli and bananas.

I ate a banana and spinach smoothie for breakfast.

Chinchillas and kittens are cute.

My sister adopted a kitten yesterday.

Look at this cute hamster munching on a piece of broccoli.

= 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)

= 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

LDA is a **generative model**

LDA represents documents as mixtures of topics that spit out words with certain probabilities. Here's how:

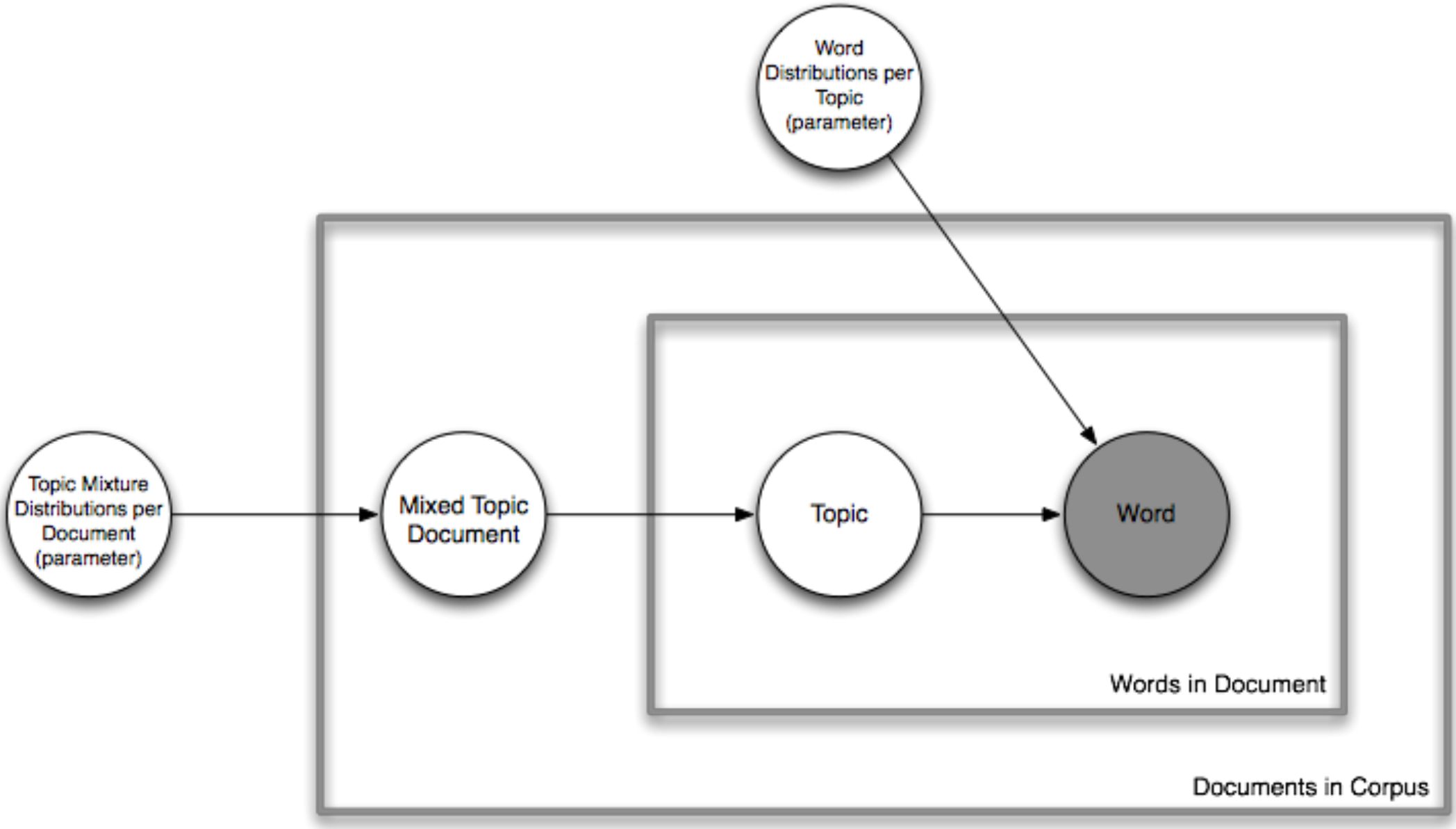
Decide on the number of words N the document will have (say, according to a Poisson distribution).

Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

Generate each word w_i in the document by:

First picking a topic: for example, you might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability).

Using the topic to generate the word itself. For example, if we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.


Word Distributions per Topic (parameter)
Topic Mixture Distributions per Document (parameter)
Mixed Topic Document
Topic
Word
Words in Document
Documents in Corpus

LDA is a **generative model**

Pick 5 to be the number of words in D.

Decide that D will be 1/2 about food and 1/2 about cute animals.

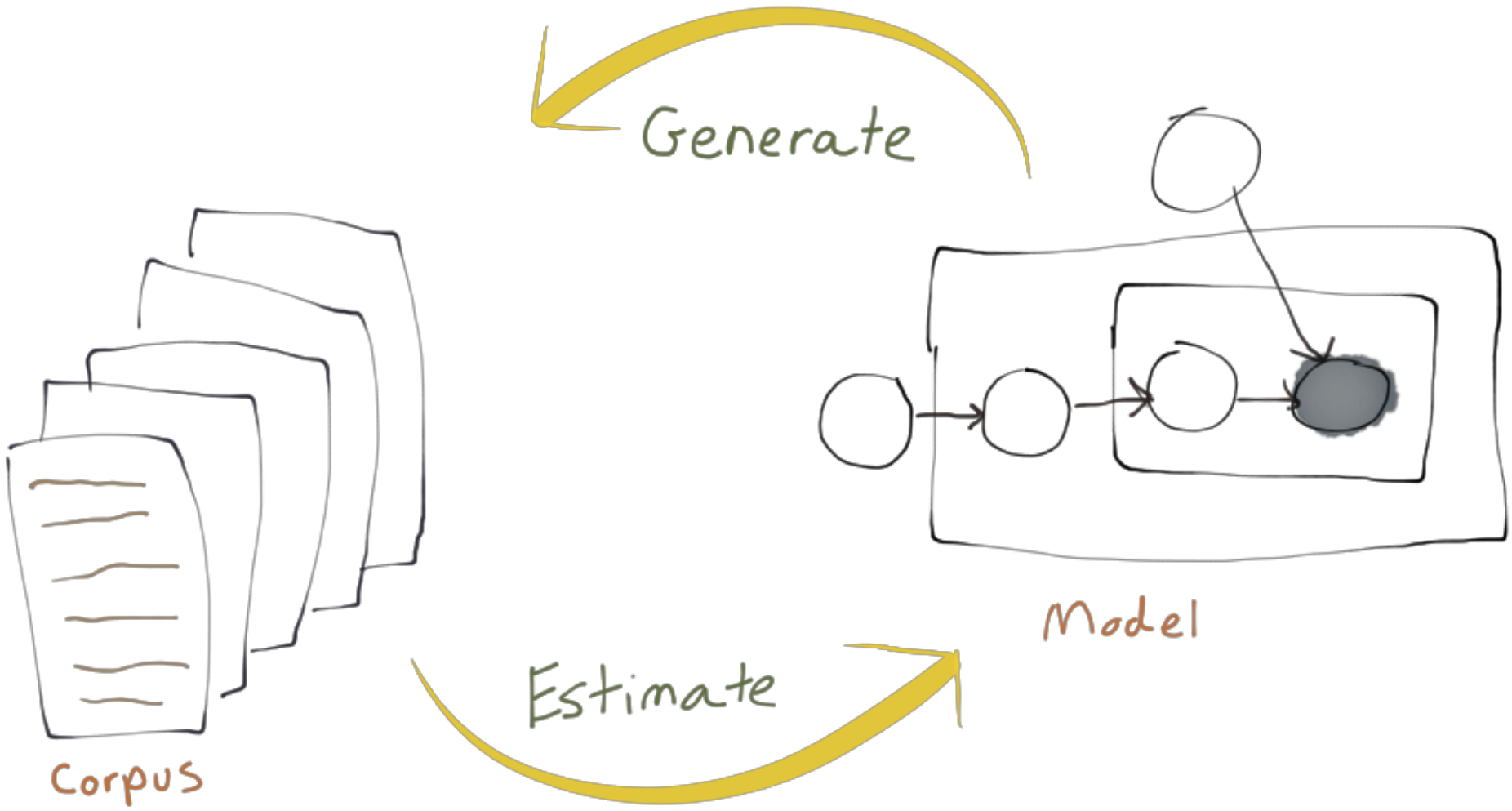
Pick the first word to come from the food topic, which then gives you the word “broccoli”.

Pick the second word to come from the cute animals topic, which gives you “panda”.

Pick the third word to come from the cute animals topic, giving you “adorable”.

Pick the fourth word to come from the food topic, giving you “cherries”.

Pick the fifth word to come from the food topic, giving you “eating”.



http://mcburton.net/blog/joy-of-tm/

Learning

Go through each document, and randomly assign each word in the document to one of the K topics.

Notice that this random assignment already gives you both topic representations of all the documents and word distributions of all the topics (albeit not very good ones).

So to improve on them, for each document d:

> Go through each word w in d:
>
> > And for each topic t, compute two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Reassign w a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t)
> >
> > In other words, in this step, we're assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.

After repeating the previous step a large number of times, you'll eventually reach a roughly steady state where your assignments are pretty good. So use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).