TOPIC MODELING THE ‘GREAT UNREAD’

Peter Leonard

Alabama Digital Humanities Center | March 2013
The Challenge of Scale

* The next 5 years will be defined by the challenge of scale

* Big Datasets
  * Google Books, HathiTrust, JSTOR DFR, Internet Archive, DIYBookScanner.org...

* Big Intellectual Approaches
  * Distant Reading, Algorithmic Criticism, Scalable Reading, Humanities Macroscopes...
The Challenge of Scale

* Can we get the benefits of algorithms without losing sight of the text?
* Can literary expertise help make sense of Big Data?
* Will Computer Science colonize the Humanities?
The majority of books disappear forever—and “majority” actually misses the point: if we set today’s canon of nineteenth-century British novels at two hundred titles (which is a very high figure), they would still be only about 0.5 percent of all published novels.

“The Slaughterhouse of Literature”
Franco Moretti
The Great Unread

Full-Text Scandinavian Books Digitized per Decade

- Danish
- Swedish
- Norwegian

Estimated as of March 2010
Google Books Digital Humanities Grants

- Collaborative research program to explore the digital humanities using the Google Books corpus.

- “Google’s first formal foray into supporting humanities text-mining research on its corpus of more than 12 million digitized books.”  
  *Chronicle of Higher Education* July 14, 2010

- 12 Teams Worldwide

- One doing Scandinavian Literature
Will the Great Unread remain so?
Nobody can read 160,000 books

* 160,000 books x 40,000 words each = 6,400,000,000.

* Redefine what it is to read.


* Move from the explicit (subject headings) to the implicit (latent variables).
What do you do with a million books?

D-Lib Magazine
March 2006

Volume 12 Number 3
ISSN 1082-9873

What Do You Do with a Million Books?
Gregory Crane
Tufts University
<gregory.crane@tufts.edu>

Introduction

The Greek historian Herodotus has the Athenian sage Solon estimate the lifetime of a human being at c. 26,250 days ([Herodotus, The Histories, 1.32](https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0156:book1.32)). If we could read a book on each of those days, it would take almost forty lifetimes to work through every volume in a single million book library. The continuous tradition of written European literature that began with the *Iliad* and *Odyssey* in the eighth century BCE is itself little more than a million days old. While libraries that contain more than one million items are not unusual, print libraries never possessed a million books of use to any one reader. The great libraries that took shape in the nineteenth and twentieth centuries were meta-structures, whose catalogues and finding aids allowed readers to create their own customized collections, building on the fixed classification schemes and disciplinary structures that took shape in the nineteenth century.

The digital libraries of the early twenty-first century can be searched and their contents transmitted around the world. They can contain time-based media, images, quantitative data, and a far richer array of content than print, with visualization technologies blurring the boundaries between library and museum. But our digital libraries remain filled with digital incunabula – digital objects whose form remains firmly rooted in traditions of print, with HTML and PDF largely mimicking the limitations of their print predecessors.
There’s an inherent (passive) relationship here. Can there be an active one?
Many Approaches

* Search for Terms
* Use Cataloging Metadata
* Let the Data Organize Itself
Many Approaches

- Search for Terms
- Use Cataloging Metadata
- Let the Data Organize Itself
Keyword Search in Philologic

1. 1884_Wild_Duck (bib:p.81) yrkisk and; for det er en vildand. Gregers. Nej, er det virkelig? En vild a
2. 1884_Wild_Duck (bib:p.82) en», som De sa’, – det er vildanden, det. Det er et vild- and, far. Hedvig
3. 1884_Wild_Duck (bib:p.82) and, far. Hedvig. Min vildand. For jeg ejer den. Gregers. Og den kan l
4. 1884_Wild_Duck (bib:p.84) en, løjtnant Ekdal. Deres vildand kom da op igen. Ekdal. Han havde slig en
5. 1884_Wild_Duck (bib:p.85) a han hørte dette her med vildanden, så maged han det så, så
6. 1884_Wild_Duck (bib:p.88) t skal gå med mig som med vildanden – Gina. Hm, jeg synes nok De skulde så
8. 1884_Wild_Duck (bib:p.112) gjort den nye kurven til vildanden. Gregers. Det kunde De, ja; og De var j
10. 1884_Wild_Duck (bib:p.112) ers. Kan tænke det; for vildanden er vel den aller fornemste der inde. He
12. 1884_Wild_Duck (bib:p.120) bordet). Den velsignede vildanden, ja. Den gører der da krusifikser nok
15. 1884_Wild_Duck (bib:p.154) lsen. Hedvig. Og så med vildanden, far, – og med alle hansene og kanin
16. 1884_Wild_Duck (bib:p.154) morgen da. Den fordomte vildanden skulde jeg helst ha’ lyst til at vri
18. 1884_Wild_Duck (bib:p.154) n, far, – det er jo min vildand! Hjalmar. Derfor gør jeg det heller ikke
19. 1884_Wild_Duck (bib:p.155) r efter ham). Men tænk, vildanden, – den skakkers
20. 1884_Wild_Duck (bib:p.156) må da endelig ikke gøre vildanden noget ondt, ondt
23. 1884_Wild_Duck (bib:p.196) ed). Ja næsten mere. Vildanden har vi jo også fået sendendes til for
24. 1884_Wild_Duck (bib:p.196) Gregers (afledende). Ja vildanden, det er sandt! Lad os snakke lidt om vi
27. 1884_Wild_Duck (bib:p.196) for jeg læser bøn for vildanden hver aften og bør, bør
28. 1884_Wild_Duck (bib:p.197) og nu bør be’r De for vildanden også? Hedvig. Jeg syntes, det var beds
29. 1884_Wild_Duck (bib:p.197) s, det var bedst at ta’ vildanden med; for hun var så sygelig i førstni
30. 1884_Wild_Duck (bib:p.198) öd for. Gregers. Og den vildanden, som De holder så inderlig af, den vil
31. 1884_Wild_Duck (bib:p.198) nu De frivillig offred vildanden for hans skyld? Hedvig
32. 1884_Wild_Duck (bib:p.198) Hedvig (rejser sig). Vildanden! Gregers. Om De nu for ham offervi
33. 1884_Wild_Duck (bib:p.199) har hvis bedstefor skyde vildanden for mig. Gregers, ja, gen så. Men ikk
Keyword Search in Google Books Ngram Viewer

Graph these case-sensitive comma-separated phrases: Sweden, Norway
between 1800 and 2000 from the corpus English with smoothing of 3.

Search lots of books

Swedish and Norwegian trends over time.
Problems with Keyword Search

* The researcher *doesn’t know what to search for.*

* Stanford project on Martha Ballard’s diary:
  * God referred to as “author of all my mercies.”


* Search Results for “vampyr”: 1

* To a first-order approximation, this book is *not* about vampires.
Many Approaches

* Search for Terms
* Use Cataloging Metadata
* Let the Data Organize Itself
Problems with Metadata

The Chronicle Review

August 31, 2009

Google's Book Search: A Disaster for Scholars

By Geoffrey Nunberg

Whether the Google books settlement passes muster with the U.S. District Court and the Justice Department, Google's book search is clearly on track to becoming the world's largest digital library. No less important, it is also almost certain to be the last one. Google's five-year head start and its relationships with libraries and publishers give it an effective monopoly: No competitor will be able to come after it on the same scale. Nor is technology going to lower the cost of entry. Scanning will always be an expensive, labor-intensive project. Of course, 50 or 100 years from now control of the collection may pass from Google to somebody else—Elsevier, Unesco, Wal-Mart. But it's safe to assume that the digitized books that scholars will be working with then will be the very same ones that are sitting on Google's servers today, augmented by the millions of titles published in the interim.
Problems with Metadata
Many Approaches

- Search for Terms
  - Presumes *a priori* knowledge of corpus
- Use Cataloging Metadata
  - Metadata isn’t always good enough
- Let the Data Organize Itself
  - Crazy Talk?
Self-Organizing Data

* Rely on statistical patterns to separate signal from noise

* One example of such a pattern: Term Co-Occurrence

  “You shall know a word by the company it keeps.”

  –J.R. Firth, 1957

* Discover topics from observed documents: Topic Modeling
Topic Modeling Visualized

Blei 2001
Self-Organizing Data: Midwife’s Diary from Maine

Mining the Dispatch

"Mining the Dispatch," seeks to explore—and encourage exploration of—the dramatic and often traumatic changes as well as the sometimes surprising continuities in the social and political life of Civil War Richmond. It uses as its evidence nearly the full run of the Richmond Daily Dispatch from the eve of Lincoln’s election in November 1860 to the evacuation of the city in April 1865. It uses as its principle methodology topic modeling, a computational, probabilistic technique to uncover categories and discover patterns in and among texts. On this site you’ll be able to view and generate graphs and charts that reveal some of the changing patterns in the topics that dominated the news during the Civil War in the capital of the Confederacy’s newspaper of record.

Topic Modeling as DH

- 2013 Special Issue of *Poetics*
  - Focused on Topic Modeling in Humanities & Social Sciences
- Folks to follow:
  - Ben Schmidt (Cultural Observatory, Harvard)
  - Ted Underwood (English, UIUC)
- November 2012 NEH Institute for Advanced Topics in DH
- Topic Modeling Workshop at MITH - University of Maryland
Three Assumptions:
✦ Documents contain Topics.
✦ Topics ‘prefer’ certain words.
✦ Documents might have been created by choosing words from Topics randomly.
"Let’s see, I want to write a book about...

Marriage

Alchemy

giftas
hustru
man
adoptivbarn
säker
guld
järnsulfat
metalliskt
kemi
tvångströjan
galen

...so I’ll just pick some words at random*
from each category and throw them together...”

* probably not how Strindberg wrote, but who knows?
Three Assumptions:
- Documents contain Topics.
- Topics ‘prefer’ certain words.
- Documents might have been created by choosing words from Topics randomly.
  (This “generative” process is imaginary but statistically sound.)

One Amazing Magic Trick:
- Hidden Structure is inferred by reversing the generative process.
There's an inherent (passive) relationship here.
Can there be an active one?

Scandinavian Studies: 
- Danish Folklore
- Stage Realism
- Hans Christian Andersen
- Novels about the countryside
- Women writers of the Modern Breakthrough

The Great Unread: 
- Scandinavian Books
- 1800 – 1923

Literary Fields vs. Google Books
One Strategy

- Topic-Model sub-corpora that are well-known and well-understood
  - Darwin’s works?
  - Ibsen Stage Realism?
  - Danish Folklore?

- Concentrate our efforts on cleaning, curating and perfecting those models

- Apply these curated topic models to Google Books

- Expose passages in books that score highly for interesting topics.
Technical Infrastructure

* Split books up into paragraphs
* Build SQL database for volume metadata
* Invoke Topic Modeling module of Machine Learning library from UMass-Amherst
* Visualize results online
* Enable researchers to label & curate topics
* Provide deep links back into Google Books reading environment
Evald Tang Kristensen

* Danish, active 1867 – 1929
* 24,000 field diary pages
* 3,500 discrete individuals
* 85 published volumes
* 13,903 tales in *Danske Sagn* alone
Imaginary Generative Process

“Let’s see, I want to tell Evald a story about...

...so I’ll just pick some words at random* from each category and throw them together...”

* probably not how people tell stories, but who knows?
Topic Modeling in Action: Danish Folklore

**Hearing Voices**

- **words:**
  - samme
  - natten
  - natten
  - hørte
  - stille
  - salmer
  - skrig
  - tydelig
  - hørte
  - lyd
  - underlig
  - hørte
  - lyd
  - tydelig
  - lyd
  - hørte

- **phrases:**
  - samme øjeblik
  - sang salmer
  - inde hos
  - ude marken
  - samme hørte
  - klagende lyd
  - tydelig hørte
  - hørte skrig
  - hørte
  - samme

**Hearing Voices Chunks per Year**

- Bar chart showing the number of chunks per year.

**Hearing Voices-related topics:**

- Topic #91
- Topic #97
- Topic #44
- Haunted Fame
- Danes/Swedes
- Dallying
- Ringing Bells
- Forests
- Life and Death

**Book Chunks 1871-1922:**

- Knud Knudsen Læsøboeg i dansk-nordisk spillegreve
  - 1880
  - Chunk #1250
- Knud Knudsen Læsøboeg i dansk-nordisk spillegreve
  - 1880
  - Chunk #1197
- Vilhelm Bergsøe Paa kale stifter.
  - folksmedværk.
  - samlede
  - 1906
  - Chunk #2504
- Henri Jorgensen Hulstof-Kaae
  - Christensens theaterhistorie
  - 1876
  - Chunk #179
- Laurids Brunn Kropsteins saga
  - 1904
  - Chunk #872
- Henri Jørgensen
  - Lykke-Per Lykke-Per og hans kæreste
  - 1902
  - Chunk #90
- Emil Aarestrup; Frederik Ludwig Lieberdine
  - Samlede digte og dragter
  - 1860
  - Chunk #3640
- J. L. J. L. Eriksen
  - Samlede værker
  - 1902
  - Chunk #124
- Holger Theodor Højmark
  - Vigtige skrækkendere
  - 1902
  - Chunk #1207
- Ottokar Gotthard Monrad
  - Flyttning fra
  - 1902
  - Chunk #1207

**Lykke-Per, Lykke-Per og hans kæreste:**

- Published 1902
- Go to Book | Find Chunk in Book

**Ved Lyden af hans Skridt vendte hun Hovedet 19 om og:**

- Standshedde ham med et Øjeblik, der ganske Yst var
  - Ulykkelig og skamfuldt, men som dog meget afgørende,
  - Ja forbindet forbed ham at komme nærmere. Og da der nu
  - Hørte flere Stemmer inde i Høvesalen, blev hun et Øjeblik
  - Ganske hvid af Angst og skyndte sig forbi ham med bejelje
  - Hoved.
For a topic we labeled “death and churchyards,” the following passage appeared as a high rank result:

Yesterday, when I saw him, I came to think — God knows how — about a starving dog, no, not starving, but a miserable, tired, emaciated dog that lies still, eyes heavy and dies on his master’s grave. And I don’t know, but now I find this picture striking: thought, the controlling, the dominant forces in him have died, and now he spiritually starves to death on his dead master’s grave.

— Herman Bang, *Haabløse Slægter* (1880) p. 319

Bang captures the uncanny, even supernatural, connection in folk belief between dogs and their masters after death. We found this not by doing a keyword search for “grave” in Bang, but rather by letting the folkloric data organize itself and attract similar passages in 1880s Danish literature.
Darwin in Denmark

1859: *On the Origin of Species by Means of Natural Selection* (trans. 1872)

1871: *The Descent of Man, and Selection in Relation to Sex* (trans. 1874-5)
JP Jacobsen: Science, Translations & Literature

- 1868: *En Arabesk*
- 1869: *Gurrerange*
- 1871: *Menneskeslægtens Oprindese*
- 1872: *Om Arternes Oprindelse ved Kvalitetsvalg eller ved de heldigst stillede Formers Sejr i Kampen for Tilværelsen*
- 1873: University of Copenhagen Gold Medal for research on freshwater algæ.
- 1874: *Arabesk. Til en Haandtegning af Michel Angelo* (Poetic Debut)
- 1874-5: *Menneskets Afstamning og Parringsvalget*
- 1875: “En skud i taagen”
- 1876: *Marie Grubbe*
- 1880: *Niels Lyhne*
- 1882: “Pesten i Bergamo”
Topic Modeling in Action: Darwinian Thought
“Social Instinct” in *Det nye aarhundrade & Inger*

In a reformist piece on the subject of “Det gældende Straffesystem” [The Current Penal System], a largely forgotten yet at the time influential Police Inspector, August Goll (1866-1936), laments the unfairness of Danish criminal law as:

“...truly a tragic conflict, in which Society as the strongest crushes the weakest, without the slightest moral right to do so—for in the zone of morality no dictate can apply” (Goll 1906, 409).

Darwin’s conception of “social instinct” had broad appeal across many fields. In Jakob Knudsen’s (1858-1917) novel *Inger*, a man asks the parish minister whether he considers his affair with Inger to be dishonorable (*æreløst*). The minister responds:

Yes, Ditlev, I must. And that is dishonorable you will notice more strongly and clearly each day it continues unfortunately. Because it is society alone that decides what is honor and what is shame. You have offended society’s morals and laws, such as they are nowadays, and that is what counts (and must count) with respect to honor and shame—no matter how good a conscience you may have had in your own ignorance (Knudsen 1906, 253).
Topic Modeling in Action: Women & the Modern Breakthrough
Review of Topic Modeling

- Textual data at medium- and large-scale can organize itself, presenting “Topic Models” as patterns of word co-occurrence.

- Scholars can use their hard-won domain expertise to evaluate these Topic Models and select some the most interesting.

- We can use these thematic Topic Models to explore both:
  - Lesser-known or forgotten passages within a sub-corpus (intrinsic)
  - The “great unread” of the 19th Century print explosion (extrinsic)
Critiques

* reductive
* mechanical
* superficial
* anti-humanistic
Algorithmic Criticism

“Any reading of a text that is not a recapitulation of that text relies on a heuristic of radical transformation... a new text in which the data has been paraphrased, elaborated, selected, truncated, and transduced.”

— Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*
Critiques

* reductive
* mechanical
* superficial
* anti-humanistic
Possibilities

* usefully reductive
* mechanical where it counts
* superficial pointers to enable deep readings
* trans-humanistic (human & machine)
The digital transformation of scholarship has brought problems of scale, but also innovation in approach.

Complementing their historic skills in close reading, humanists may want to explore letting data organize itself through algorithms.

Done correctly, this may lead to more close reading — at various scales — of more kinds of texts.
Antager man Afstamningslæren, saa vil Mennesket ... ikke længere betragte sig som en Undtagelse fra Naturlovene, men vil endog begynde at se efter det lovmæssige i sine egne Handlinger og Tanker och stræbe efter at faa sit eget Liv i Overensstemmese met Naturlovene.

If one accepts the teachings of evolution... then Man will no longer regard himself as an exception from the laws of nature, but will begin notice these rules in his own actions and thoughts, and strive to place his own life in congruence with the laws of nature.

English-Language Demo
"I want to write a poem about...

Industry
Packing Tool
Railroads
Brutal Factory

Masculinity
Fierce Husky Brutal Brawling

Chicago
Crooked City Gunmen

...so I’ll just pick some words at random* from each category and throw them together!"

* probably not how Sandburg composed poetry, but who knows?
“Lead Soldiers”
Amy Lowell
1915

LEAD SOLDIERS HE nursery fire burns brightly, crackling in cheerful little explosions and trails of sparks, up the back of the chimney. Miniature rockets peppering the black bricks with golden stars, as though a gala flamed a night of victorious wars. The nodding mandarin on the bookcase moves his head forward and back, slowly, and looks into the air with his blue-green eyes. He stares into the air and nods-forward and back. The red rose in his hand is a crimson splash on his yellow coat. Forward and back, and his blue-green eyes stare into the air, and he nods-nods. Tommy’s soldiers march to battle, Trumpets flare and snare-drums rattle. Bayonets flash, and sabres glance [269] How the horses snort and prance! Cannon drawn up in a line Glitter in the dizzy shine Of the morning sunlight. Flags Ripple colors in great jags. Red blows out, then blue, then green; Then all three—a weaving sheen Of prismatic patriotism. March Tommy’s soldiers, stiff and starched, Boldly stepping to the rattle Of the drums, they go to...
**Symphony of a Mexican Garden**

Grace Hazard Conkling

1912

SYMPHONY OF A MEXICAN GARDEN

1. THE GARDEN Poco sostenuto in A major. The laying tide of inarticulate air. Fivace in A major. The iris people dance. 2. THE POOL Allegretto in A minor Cool-hearted dim familiar of the doves. 3. THE BIRDS Presto in F major I keep a frequent tryst. Presto meno assai. The blossom-powdered orange-tree. 4. TO THE MOON Allegro con brio in A major. Moon that shone on Babylon. TO MOZART. What junipers are these, inlaid with flame of the pomegranate tree? The god of gardens must have made them. This still unrumored place for thee. To rest from immortality. And dream within the splendid shade. Some more elusive symphony Than orchestra has ever played.

11. I In A major Poco sostenuto. The laying tide of inarticulate air. Breaks here in flowers as the sea in foam. But with no satirist of falling wave. The odor-laden winds are very still. An unimagined music here exhaled in upcurled petal, dreamy bud half-furled, and variations of thin vivid leaf. Symphonie beauty that some god forgot. If form could waken into lyric sound. This flock of irises like poising birds would feel song at their slender feathered throats. And pour into a grey-winged ari. The wrinkled silver fingermarked with pearl. That flight of ivory roses high along. The airy azure of the larkspur spires. Would be a fugue to puzzle nightingales. With too-evasive rapture, phrase on phrase. Where the hibiscus flares would cymbals clash. And the black cypress like a deep bassoon. Would hum a clouded amber melody. But all across the trudging ragged chords. That are the tangled grasses in the heat. The mariposa ilies fluttering. Like trills upon some archangelic flute. [12] Symphony of a Mexican Garden

Grace Hazard Conkling

1912
THE INTANGIBLE SYMPHONY

How shall I capture Sound and desire? Let candor stir upon candor As sword upon sword, Tempering the tenor and the timbre Of this’ sweet ecstasy. Grieved is my mind, Harassed by music Untouched of any sound. Yet on trelle, on infinite arch, On bridges of fretted iron Frail to thought, acrid to sight, Thunderous with traffic of men Red-budding, peach-petaled Beauty flames into view. But how shall I capture Sound and desire? How shall I hear The pointed vagaries, The evanescent harmonies, That float unfingered Across the strings of the mind? [136] Mark Turbyfill How shall I hear, Plucked from the intangible mind-strings, The song desire sings, and sings? There is no create instrument.