Exploring a digital library through Key Ideas

Bill Schilit & Okan Kolak
Google Research
Popular Passages within books

Saturday at 10:50 am
Popular Passages across books

```
We know now that a text consists not of a line of words, releasing a single “theological” meaning (the “message” of the Author-God), but of a multi-dimensional space in which are married and contested several writings, none of which is original..."

The Rustle of Language - Page 52
by Roland Barthes - 1989
Limited preview - About this book

Theories of Authorship: A Reader - Page 211
by John Caughie - Performing Arts - 1981

...itself, language which ceaselessly calls into question all origins. We know now that a text is not a line of words releasing a single “theological” meaning (the “message” of the Author-God) but a multi-dimensional space in which a variety of writings, none of them original, blend and clash. The...
Limited preview - About this book - Add to my library

On Deconstruction: theory and criticism after structuralism - Page 32
by Jonathan Culler - Literary Criticism - 1983 - 312 pages

...“we know,” write Barthes with that assurance that comes upon some writers in Paris, “that the text is not a line of words releasing a single “theological” meaning (the “message” of an Author-God) but a multi-dimensional space in which a variety of writings, none of them original...
Limited preview - About this book - Add to my library

The Ascetic Imperative in Culture and Criticism - Page 123
by Geoffrey Galt Harpham - Philosophy - 1987

... (“From Work to Text” 160. “We now know,” he says in “The Death of the Author,” “that a text is not a line of words releasing a single “theological” meaning (the “message” of the Author-God) but a multi-dimensional space in which a variety of writings, none of them original, blend and clash” (148)....
Limited preview - About this book - Add to my library
```

World’s largest quotation database
ideas ⊆ quotations

“sustainable development does imply limits - not absolute limits but limitations imposed by the present state of technology and social organization on environmental resources and by the ability of the biosphere to absorb the effects of human activities.”

– The Brundtland Report
ideas and their connections
Learning transforms us, that does which all nourishment does which does not merely "preserve"—as the physiologist knows. But at the bottom of us, "right down deep," there's, to be sure, something unteachable, a grain of eternal spiritual fact, of preordained decision and answer to preordained selected questions. In the case of every cardinal problem there speaks an unchangeable "this is it," about man and woman, for example, a thinker cannot retreat but only learn fully—only discover all that is "firm and settled" within him on this subject. One sometimes comes upon certain solutions to problems which inspire strong belief in us; perhaps one then forthwith calls them one's "convictions." Later—one sees them only as footnotes to self-knowledge, signposts to the problem which we are—more correctly, the great stumbling block which we are, to our spiritual fate, to the unattainable "right down deep."—Having just paid myself such a deal of pretty compliments I may perhaps be more ready permitted to utter a few truths about "woman as such": assuming it is now understood from the outset how great an extent these are only—my truths. —

Woman wants to be independent and to do that end she is beginning to enlighten men about "woman as such"—this is one of the world developments in the general legalization of women.
Problems in Mining and linking ideas

1. Distinguish “ideas” from other words, but at the surface all words look alike

2. No common classification (e.g., labeling) so relations are hidden
   • “Bovine Spongiform Encephalopathy (BSE)” also known as “Mad Cow Disease”
The Great Ideas

- University of Chicago & Encyclopedia Brittanica published Great Books of the Western World in 1952
- 443 works in 54 volumes
- Bacon, Hobbes, Locke, Adam Smith, Swift, ...
Syntopicon: “Angel” to “World”

2 volumes, 102 ideas, 7 years, 50 indexers, 75 clerical staff, $1 million
Great Ideas Critique

“In times like ours it is plainly impossible for 10 or 12 men, working in one single tradition, to select and to explain all the greatest books...”

- NYT Book Review, Sept 14, 1952

• Missing Equality, Civil Rights, ...

• “Not our great ideas...”
Not my books...
Our approach

1. quotations reflect popular ideas...
2. popular terms near quotes are good labels
Without editors, ideas emerge from large digital collections.
Without editors, ideas emerge from large digital collections.
Key Terms from context

Longest Quotation Text

Although we may never know with complete certainty the identity of the winner of this year’s presidential election, the identity of the loser is perfectly clear. It is the nation’s confidence in the judge as an impartial guardian of the law.

Instance Text with Pre and Post Context

1. As Justice John Paul Stevens wrote in a blistering dissent, “Although we may never know... law” And that confidence may be further shaken if independent full manual recounts by the press (allowed under Florida’s sunshine laws) decide that Mr Gore did indeed ”win”
2. It chastised the Court for holding the justices of the Florida Supreme Court up to ridicule. “Although we may never know... law”
3. “The identity of the loser is perfectly clear” wrote Justice Stevens, “It is... law.”
4. ...
5. On December 13, Al Gore conceded the election to George W. Bush. The high court’s ruling was uncharacteristically long-winded and excruciatingly tortured in its logic. Dissenting from the majority, Associate Justice Stephen Breyer wrote that while “we may never know...”

A quote from Justice Steven's opinion (Bush v. Gore Supreme Court Case) with 81 instances
What is context?

- Goal: capture attribution and concept labels
- Few hundred bytes before and after
  - Quote with 81 instances $\rightarrow$ 1,000s words
  - diminishing returns with more context
- Look for frequent terms over all instances
- Separating the quote from the context is hard (see paper)
Two types of key terms

- **Key Terms for a quote**
  - based on occurrences across all instances in a single quotation
  - Navigate to other quotes
  - A particular quote: “Dan Quayle” with “Jack Kennedy”

- **Related terms for key terms**
  - based on co-occurrence across all quotations with a term
  - Navigate to other quotes
  - A relation: “space race” with “cold war” related
Both quotes and key terms exhibit qualities of smart crowds

“It needs to be diverse, so that people are bringing different pieces of information to the table. It needs to be decentralized, so that no one at the top is dictating the crowd's answer. It needs a way of summarizing people's opinions into one collective verdict. And the people in the crowd need to be independent, so that they pay attention mostly to their own information, and not worrying about what everyone around them thinks.”

- James Surowiecki in Wisdom of the Crowds
Future work: idea networks

Diagram showing connections between terms such as Anthropology, Cultural Anthropology, Stanley Milgram, Holocaust, Sociology, Psychology, Erving Goffman, Sociocultural Systems, Feminism, Organizational Behavior, Groupthink, Social Psychology, and Ethnography.
Future work

• How have ideas caught on or evolved?

• Use over time and place

• Which ideas lost popularity after the demise of the Ottoman Empire?

• How long did it take Charles Darwin's ideas about evolution to influence various cultures?
Related work

Transclusion in Ted Nelson’s Xanadu
Conclusion

- Experimental but compelling aspects
- million+ books exposes lots of ideas
- navigate from idea to idea via key terms
- Crowd wisdom complements collective intelligence
- Key Ideas - skim ideas and read deeper in sources
- Wikipedia - overview & how ideas fit together
questions?

• “Generating links by mining quotations,” presentation this Saturday at 10:50 am

• Google is hiring student interns.