

Big Data Case Studies

Sayan Bhattacharyya

The HathiTrust Research Center's Extracted Features Dataset:

An Opportunity for "Distant" Reading of Millions of Books from the World's Great Research Libraries

- Sayan Bhattacharyya (<u>sayan@illinois.edu</u>)
- Peter Organisciak (organis2@illinois.edu)
- J. Stephen Downie (Project PI)
 Graduate School of Library and Information Science, UIUC, Urbana-Champaign
- Loretta Auvil and Boris Capitanu Ted Underwood
 (Illinois Informatics Institute, UIUC) (Department of English, UIUC, Urbana-Champaign)

WHAT

is

IT?

a dataset of page-level extracted features for scanned books in the HathiTrust Digital Library

Raw Text



Translation into features

(we drop you off here)



Algorithmic Use

had taken mental note of everything that was on that table. There were three plates laid, so that Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so that the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly.
"This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something that for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck agressively through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid you weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips twitched understandingly. She had expected Mrs. Rachel up; she had known that the

"that": { "DT": 1, "IN": 4, "WDT": 1 "the": { "DT": 5 "there": { "EX": 2 "thin": { "JJ": 1 "this": { "DT": 1 "though": { "IN": 1 "thought": { "VBD": 1 "three": { "CD": 1 "through": { "IN": 1 "to": { "TO": 2 "to-day": { "NN": 1 "twisted": { "VBN": 1 "twitched": { "VBD": 1 Digitized by Google

(Determiner, Preposition or subordinating conjunction, Wh-determiner)

(Verb, past tense)

as a dark and stormy night; the rain fell in t ot at occasion gust of wind that our sce agitating th gainst the di s of London nen of the po nding his so rent shops at he appearance d,—and tend did not seem eived were couched in the negative;

"Say, ye opprest by some fantastic woes, Some jarring nerve that baffles your repose, Who press the downy couch while slaves advance With timid eye to read the distant glance; Who with sad prayers the weary doctor tease To name the nameless ever-new disease: Who with mock patience dire complaints endure, Which real pain and that alone can cure: How would you bear in real pain to lie Despised, neglected, left alone to die How would ye bear to draw your latest breath Where all that's wretched paves the way to death?"

It was a dark and stormy night; the rain fell in torrents except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness. Through one of the obscurest quarters of London, and among haunts little loved by the gentlemen of the police, a man, evidently of the lowest orders, was wending his solitary way. He stopped twice or thrice at different shops and houses of a description correspondent with the appearance of the quartier in which they were situated,—and tended inquiry for some article or another which did not seem easily to be met with. All the answers he received were couched in the negative; and as he turned from each door he muttered to himself, in no very elegant phraseology, his disappointment and discontent. At length, at one house, the landlord, a sturdy butcher, after rendering the same reply the inquirer had hitherto received, added,-"But if this vill do as vell, Dummie, it is quite at your Dansing reflectively for a moment Dummia re-

ach door he muttered to himself, in no very

treets (for the houseto lamps that of the ob ittle loved the lowest d twice or tion corresp which the article or a

All the a

and as he

was checke

had taken mental note of everything that was on that table. There were three plates laid, so that Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so that the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly.

"This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something that for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck agressively through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid you weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips twitched understandingly. She had expected Mrs. Rachel up; she had known that the

sight of Matthew jaunting off so unaccountably would be too much for her neighbour's curiosity.

"Oh, no, I'm quite well although I had a bad headache yesterday," she said. "Matthew went to Bright River. We're getting a little boy from an orphan asylum in Nova Scotia and he's coming on the train to-night."

If Marilla had said that Matthew had gone to Bright River to meet a kangaroo from Australia Mrs. Rachel could not have been more astonished. She was actually stricken dumb for five seconds. It was unsupposable that Marilla was making fun of her, but Mrs. Rachel was almost forced to suppose it.

"Are you in earnest, Marilla?" she demanded when voice returned to her.

"Yes, of course," said Marilla, as if getting boys from orphan asylums in Nova Scotia were part of the usual spring work on any well-regulated Avonlea farm instead of being an unheard of innovation.

Mrs. Rachel felt that she had received a severe mental jolt. She thought in exclamation points. A boy! Marilla and Matthew Cuthbert of all people adopting a boy! From an orphan asylum! Well, the world was certainly turning upside down! She would be surprised at nothing after this! Nothing!

"What on earth put such a notion into your head?" she demanded disapprovingly.

This had been done without her advice being asked, and must perforce be disapproved.

"Well, we've been thinking about it for some time—all winter in fact," returned Marilla. "Mrs. Alexander Spencer was up here one day before

had taken mental note of everything that was on that table. There were three plates laid, so that Marilla must be expecting some one home with Matthew to tea; but the dishes were every-day dishes and there was only crab-apple preserves and one kind of cake, so that the expected company could not be any particular company. Yet what of Matthew's white collar and the sorrel mare? Mrs. Rachel was getting fairly dizzy with this unusual mystery about quiet, unmysterious Green Gables.

"Good evening, Rachel," Marilla said briskly.
"This is a real fine evening, isn't it? Won't you sit down? How are all your folks?"

Something that for lack of any other name might be called friendship existed and always had existed between Marilla Cuthbert and Mrs. Rachel, in spite of—or perhaps because of—their dissimilarity.

Marilla was a tall, thin woman, with angles and without curves; her dark hair showed some gray streaks and was always twisted up in a hard little knot behind with two wire hairpins stuck agressively through it. She looked like a woman of narrow experience and rigid conscience, which she was; but there was a saving something about her mouth which, if it had been ever so slightly developed, might have been considered indicative of a sense of humour.

"We're all pretty well," said Mrs. Rachel. "I was kind of afraid you weren't, though, when I saw Matthew starting off to-day. I thought maybe he was going to the doctor's."

Marilla's lips twitched understandingly. She had expected Mrs. Rachel up; she had known that the

sight of Matthew jaunting off so unaccountably would be too much for her neighbour's curiosity.

"Oh, no, I'm quite well although I had a bad headache yesterday," she said. "Matthew went to Bright River. We're getting a little boy from an orphan asylum in Nova Scotia and he's coming on the train to-night."

If Marilla had said that Matthew had gone to Bright River to meet a kangaroo from Australia Mrs. Rachel could not have been more astonished. She was actually stricken dumb for five seconds. It was unsupposable that Marilla was making fun of her, but Mrs. Rachel was almost forced to suppose it.

"Are you in earnest, Marilla?" she demanded when voice returned to her.

"Yes, of course," said Marilla, as if getting boys from orphan asylums in Nova Scotia were part of the usual spring work on any well-regulated Avonlea farm instead of being an unheard of innovation.

Mrs. Rachel felt that she had received a severe mental jolt. She thought in exclamation points. A boy! Marilla and Matthew Cuthbert of all people adopting a boy! From an orphan asylum! Well, the world was certainly turning upside down! She would be surprised at nothing after this! Nothing!

"What on earth put such a notion into your head?" she demanded disapprovingly.

This had been done without her advice being asked, and must perforce be disapproved.

"Well, we've been thinking about it for some time—all winter in fact," returned Marilla. "Mrs. Alexander Spencer was up here one day before

WHY

should you

CARE?

- 1) It's huge
- 2) It's accessible
- 3) You can do cool stuff with it

1) It's Huge

The HathiTrust Digital Library (HTDL)

- Approximately 14 million books
 - from the world's great research libraries:
 - a large chunk of mankind's historical textual record of culture
- 1.8 billion pages
- 610 billion words
- Approximately 4.8 million of the 14 million books are in the public domain
 - Current applications are set up to work with these 4.8 million books



2) It's Accessible

One of the largest archives of pre-digital human creation, downloadable

https://sharc.hathitrust.org/features

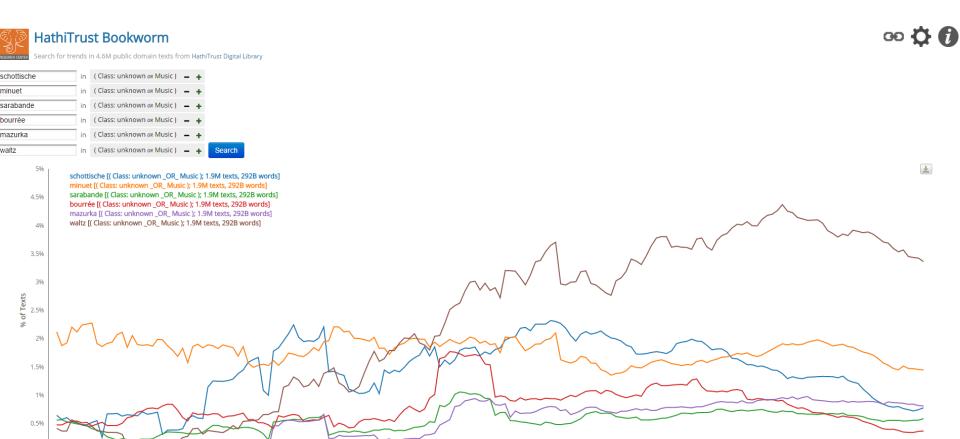
Can't share 10 million in-copyright works, but...

3) can do useful stuff with it

Large corpora allow for

- -historical
- -cultural
- -linguistic
- insights

1817 logic	proficiency	learned	studies	divinity	studied	philosophy	lectures	eminent	physic	students	scholar	grammar	school	latin	university	
1818 logic	faculty	philosophy	philosophical	professor	enquiries	treatise	science	study	accurate	philosopher	perspicuity	mathematical	learned	mathematics	knowledge	į
1819 logic	error	notion	have	rules	proves	language	has	treating	see	refer	nature	knowledge	grammar	correct	agree	
1820 logic	proves	points	arguments	touching	truth	sense	knowledge	universal	errors	says	matters	believe	reject	practice	assertion	č
1821 logic	are	gives	general	careful	relates	reasoning	had	conclusion	short	local	instances	ignorance	favour	were	subject	
1822 logic	philosophy	sci nce	coning	y and	Cur	rar	er on nr	philo pha	a ar or	q p.dy	faculty	philosophical	observation	faculties	knowledge	į
1823 logic	philosophy	Hatarar	Sciences	unstouc	WIIICIS	Various	101160060	imagination		characters	philosopher	modes	knowledge	reasoning	ideas	
1824 logic	philosophy	dispute	granimai	philosopher	phy s learning	irr	study	writings	rhetoric	languages	language	knowledge	intimate	expressing	aristotle	
1825 logic	knowledge	enlightened	tendency	ascrived	learning	character	principles	make	good	credit	whole	tend	taught	serve	rule	
1826 logic	natural	sciences	science	philosophy	knowledgo	ovestigation		work	universal	bservati/n	ossay	ctudy	learning	improvement	-disifty	, 1
1827 logic	writers	philotophy			iter		Giguage C	ノGW L	ノロレ	systems	merement		occavation	Nge 🛶	emonecus	
1828 logic	learning	reader	origin	admit	knowledge	absurd	remark	learned	furnish	various	view	such	natural	highest	writers	1
1829 logic	philosophy	natural	philosophical	speculations	judicious	objects	various	has	place	other	form	philosopher	most	knowledge	body	
1830 logic	philosophy	studes	sciences	study	science	learning	studied	judicious	rhetoric	philosopher	objects	notions	language	ideas	knowledge	
1831 logic	philosophy	knowledge			e _{atur} f C	study		Cicione	eminent	absurd	principles	objects	important	experience	society	
1832 logic	philosophy	theo ogy	science	neory	knowledge	philosophical	comions	theories	essay	author	author	controversy	metaphysics	cutical	a/guments	
1833 logic	mere	m nd	nti ns /	philosoph <i>y</i>	knowledge knowledge	pinciples	argument	Gran.	aburd	rason ng	U tels	Lifaculy V	Letor.	Selection	ignorance	
1834 logic	scriptures						deny	quotations	oral	priests	doctrine	controversy	fathers	quote	books	
1835 logic	philosophy	speculation	porbles	Jol.	notestant Pir t ism	atural	absurd	possess	language	strongest	learning	mind	world	peculiar	nature	(
1836 logic	philosophy	natural	reject	principles	writers	objects	minds	knowledge	remark	ingenious	system	reality	error	argument	are	
1837 logic	learning	divinity	learned	study	church	argument	locke	christian	philosophy	discourses	authentic	admitted	scholar	religion	proves	
1838 logic	study	philo soph y	universal	knowledge	consistent	systems	practical	natural	are	education	principles	illustrate	sciences	science	learning	
1839 logic	philosophy	meta _l hysical	X sierte C	ersorire	ronsistent Tory	abrano	njeraj hysic	arguments	writers	practical	demonstrate	authors	aristotle	writer	philosophical	
1840 logic	science	reasoning	philosophy	speculations	theology	clearness	aristotle	study	essay	theory	proves	metaphysical	philosophic	natural	minds	CO
1841 logic	philosophy	na ura	language	/ N/V	theology	so le y	benevole t	_ (ue)	Cryada	kno vledge		Applent)	U Olah	.or.stant	such	
1842 logic	writer	reader	phnosophical	philosophy	ignorance	arguments	origin	assumed	admirable	absurd	WORK	reference	notions	language	serious	
1843 logic	knowledge	natural	admit	adduced	points	intelligible	facts	absurd	theory	reasoning	instance	inquiry	compare	regard	made	ex
1844 logic	absurd	prejudice	reasoning	merits	knowledge	notions	regard	reasonings	opinions	nature	understanding	study	respect	infer	human	CO
1845 logic	reasoning	principles	absurd	arguments	understanding	abstract	reject	regard	knowledge	clearness	universal	morals	merits	errors	argument	
1846 logic	reasoning	natural	mere	admit	reader	sentiment	possess	argument	true	plausible	infer	logical	indication	illustration	remark	p
1847 logic	philosophy	knowledge	absurd	natural	minds	learned	ignorance	arguments	study	reasoning	universal	systems	mind	language	ingenious	
1848 logic	philosophy	philosophical	logical	ideas	conclusions	abstract	systems	aristotle	knowledge	faculty	speculative	natural	mental	notion	mere	
1849 logic	reasoning	logical	philosophy	conclusions	systems	abstract	theory	exist	essential	supposition	nature	natural	speculative	metaphysical	intelligible	
1850 logic	reasoning	erroneous	philosophy	correct	writer	language	conclusions	notions	remarks	abstract	speculations	merits	facts	arguments	writers	
1851 logic	philosophy	natural	minds	elements	comprehensive	characteristics	objects	mental	knowledge	principles	philosophical	clearness	universal	science	logical	CC
1852 logic	natural	character	mere	has	view	such	regard	possess	knowledge	experience	existence	essential	remark	peculiar	intellectual	



Bookworm

http://bookworm.htrc.illinois.edu

Faceted visualization of trends over 4.8 million books

Year

The HathiTrust+Bookworm project

Team Members:

– Current:

J. Stephen Downie, University of Illinois, Urbana-Champaign Erez Lieberman Aiden, Baylor College of Medicine Benjamin Schmidt, Northeastern University Robert McDonald, Indiana University Loretta Auvil, University of Illinois, Urbana-Champaign Peter Organisciak, University of Illinois, Urbana-Champaign Muhammad Shamim, Baylor College of Medicine Sayan Bhattacharyya, University of Illinois, Urbana-Champaign Leena Unnikrishnan, Indiana University







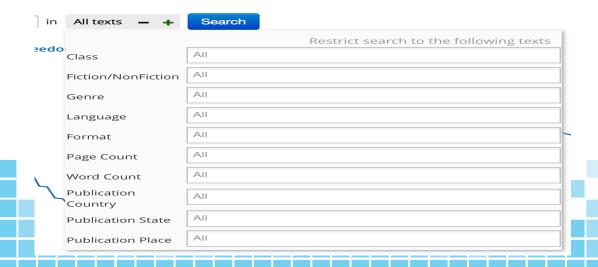
– Past:

Colleen Fallaw, University of Illinois, Urbana-Champaign Matt Nicklay, Baylor College of Medicine

Funded by an NEH Implementation Grant (2014-2016)

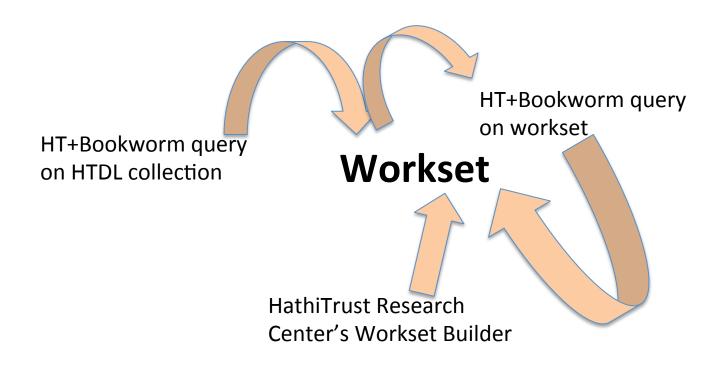
Hooking up Extracted Unigrams with Bookworm: Advantages? First advantage: Good metadata!

- HTDL has good and detailed metadata
 - metadata was meticulously created by librarians from contributing libraries
 - > allows for highly faceted queries:



Hooking up HTDL with Bookworm: Advantages?

Second advantage: HTDL's workset functionality (contd.)



Workset creation and refinement workflow

How scientific inquiry meets humanistic inquiry in culturomics as performed by HT +Bookworm

- Scientific inquiry concerns:
 - Generalization across entities
 - Discovery of patterns across entities
- Humanistic inquiry concerns:
 - Close engagement with specific entities
 - Attending to singular instances among entities

DATASET

https://sharc.hathitrust.org/features

ACKNOWLEDGEMENTS
Boris Capitanu Ted Underwood

Loretta Auvil

Colleen Fallaw J. Stephen Downie

Benjamin Schmidt (Bookworm)

Special thanks to the National Center for Supercomputing Applications (NCSA)

National Endowment for the Humanities (NEH)

Contact:
Peter Organisciak
organis2@illinois.edu

Sayan Bhattacharyya sayan@illinois.edu