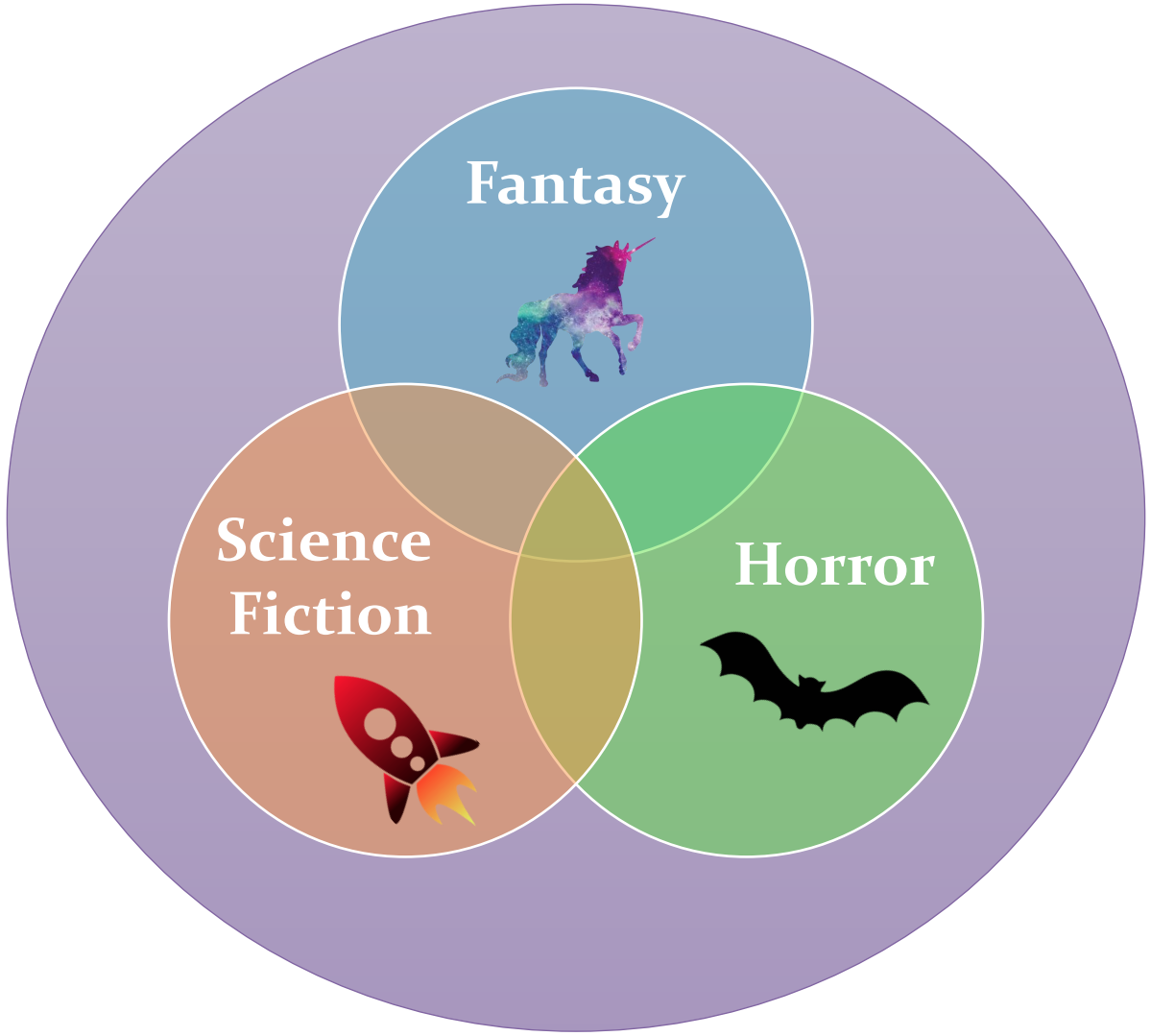
Finding Speculative Fiction in HathiTrust

Laure Thompson
UMass Amherst
@thompson_laure

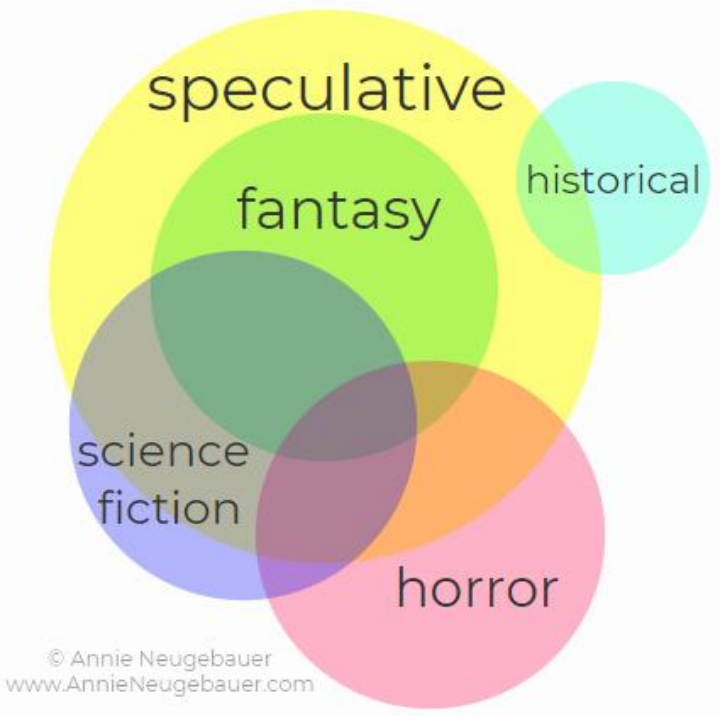
&

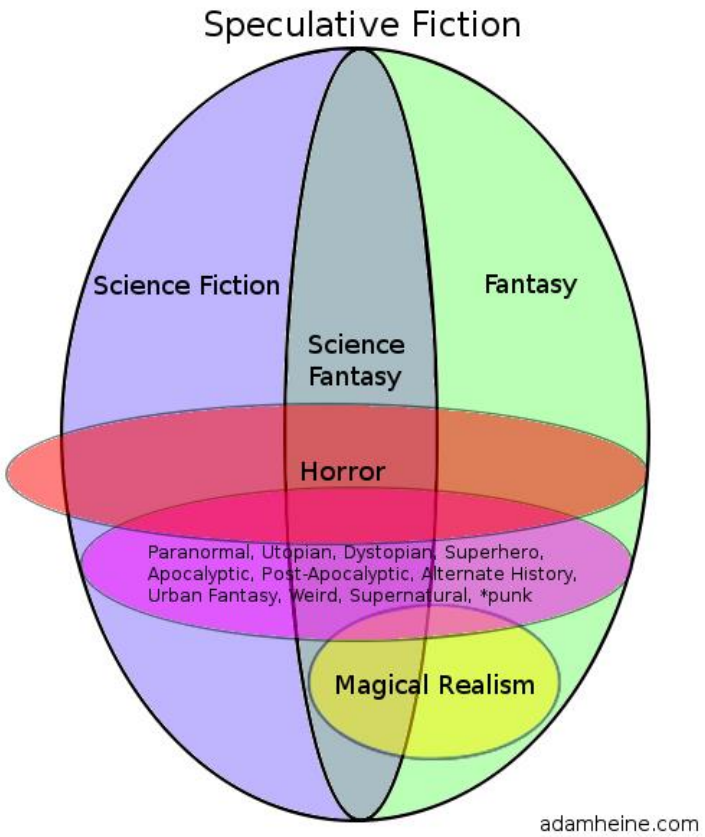
David Mimno
Cornell University
@dmimno

What is Speculative Fiction?



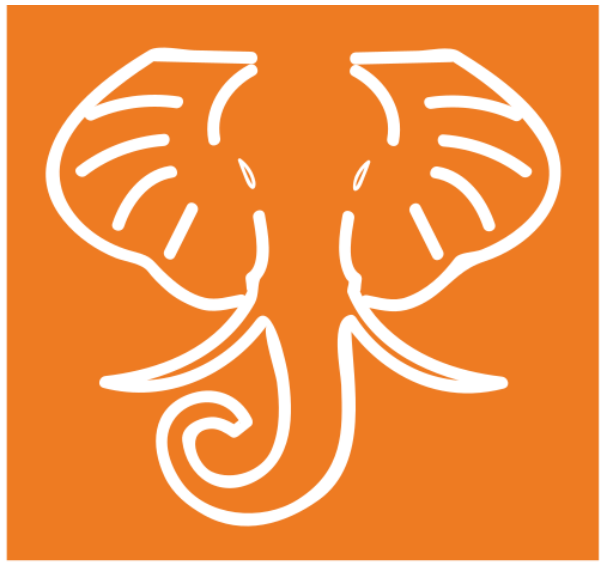
What is Speculative Fiction?





World Weaver Press

HathiTrust Digital Library



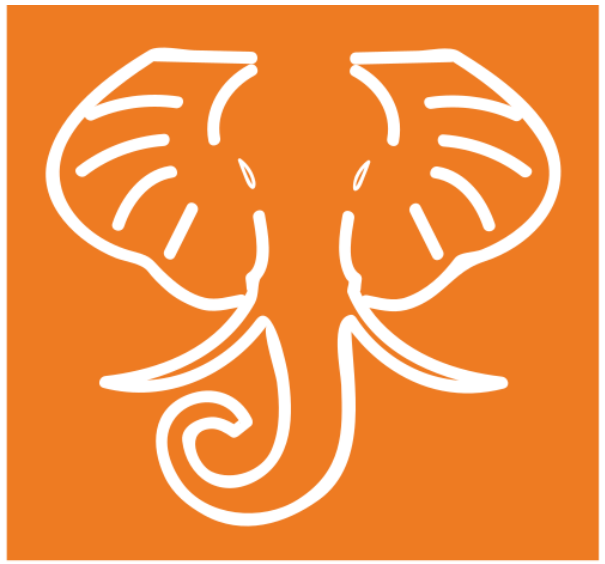
HATHI TRUST

17+ million digitized volumes

Public domain works are **viewable**

In-copyright works are **searchable**

Working with in-copyright materials



HATHI TRUST

I. HTRC Data Capsules

analytics.hathitrust.org/staticcapsules

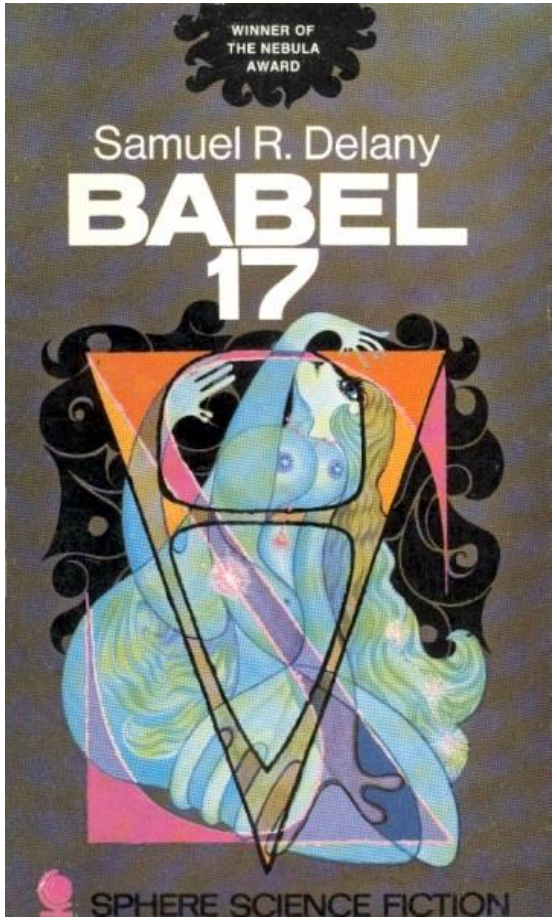
II. HTRC Extracted Features Dataset

analytics.hathitrust.org/datasets

III. HTRC Algorithms

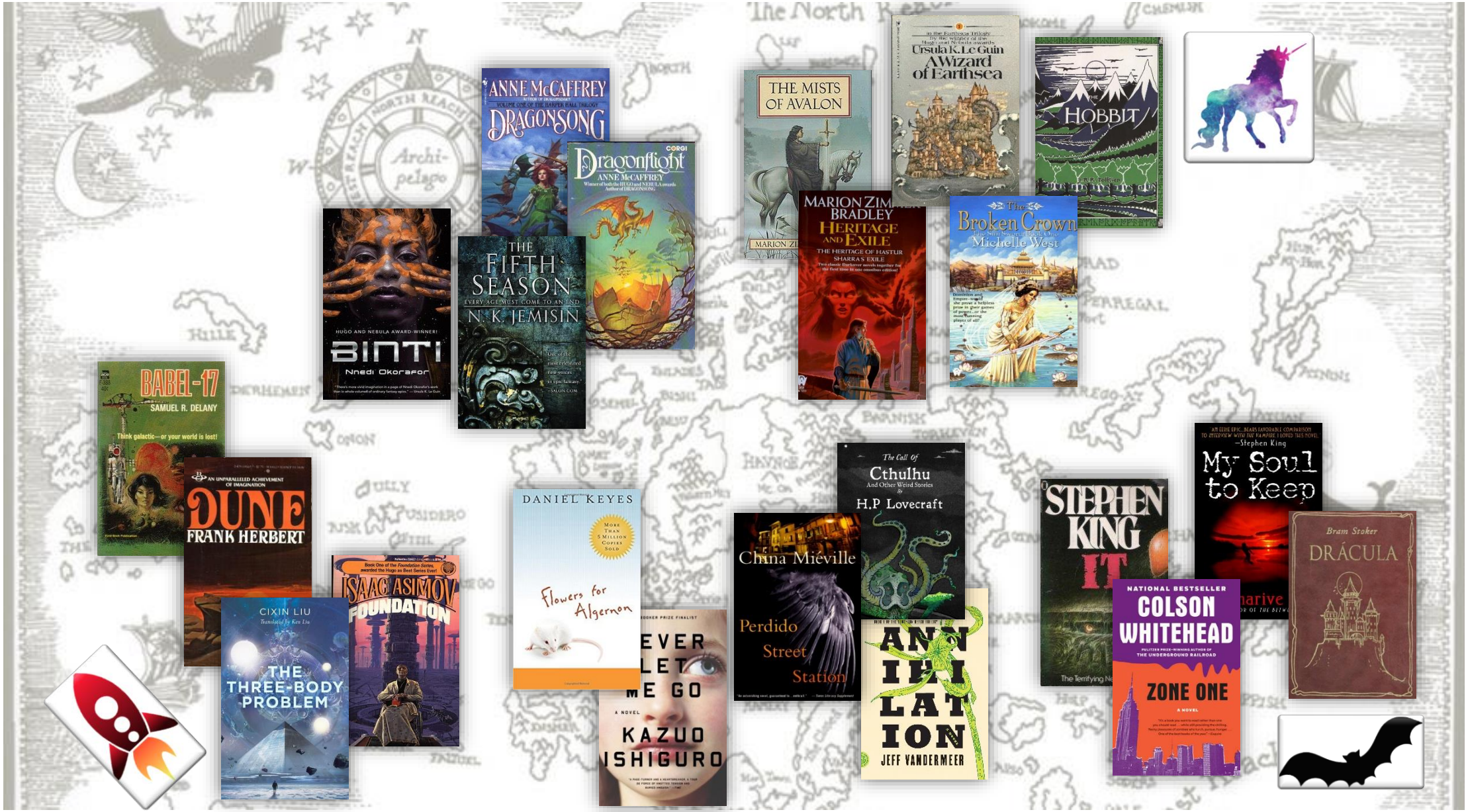
analytics.hathitrust.org/statisticalalgorithms

Extracted Features: page-level word counts

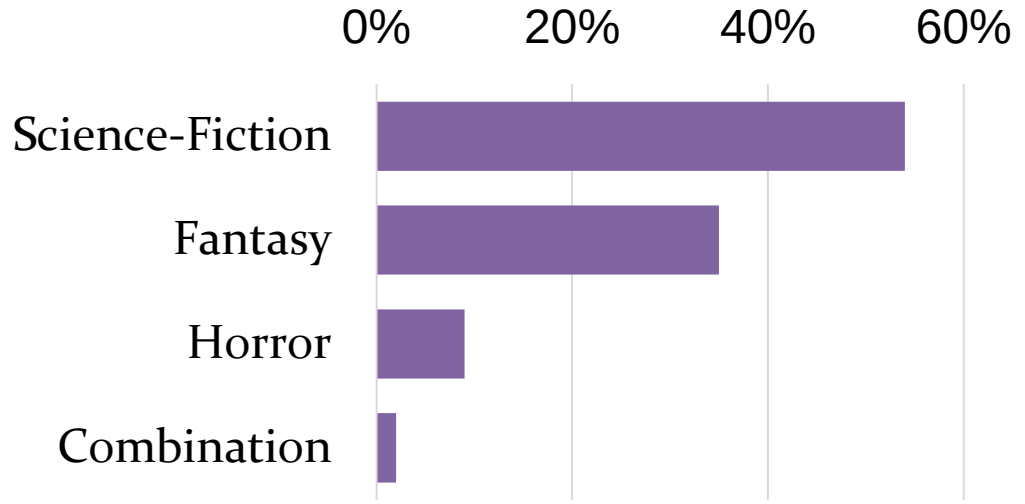


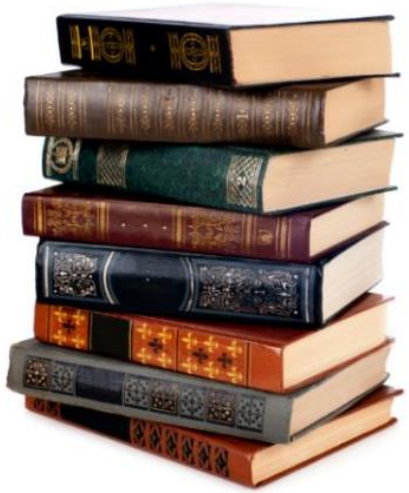
Page 77:

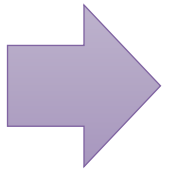
```
body: {rate: {NN:2},
       used : {VBD:1, VBN:1},
       Still : {RB:1},
       Does : {VBZ:1},
       Gorgeously : {RB:1},
       faults : {NNS: 1},
        ⋮
      }
```

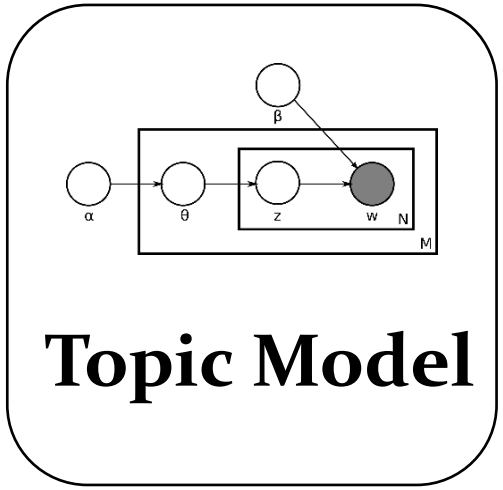

ANNE McCAFFREY
DRAGONSONG
Dragonflight
ANNE McCAFFREY
THE MISTS OF AVALON
Ursula K. Le Guin
A Wizard of Earthsea
THE HOBBIT
BINTI
Nnedi Okorafor
THE FIFTH SEASON
N. K. JEMISIN
MARION ZIMMER BRADLEY
HERITAGE AND EXILE
THE HERITAGE OF HASTUR
SHARRA'S EXILE
The Broken Crown
Michelle West
BABEL-17
SAMUEL R. DELANY
DUNE
FRANK HERBERT
ISAAC ASIMOV
FOUNDATION
CIXIN LIU
THE THREE-BODY PROBLEM
DANIEL KEYES
Flowers for Algernon
KAZUO ISHIGURO
China Miéville
Perdido Street Station
The Call Of Cthulhu
And Other Weird Stories
H.P Lovecraft
ANNIHILATION
JEFF VANDERMEER
STEPHEN KING
IT
My Soul to Keep
Stephen King
Bram Stoker
DRÁCULA
NATIONAL BESTSELLER
COLSON WHITEHEAD
ZONE ONE
A NOVEL

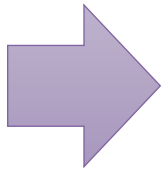
2334 novels by 903 authors








β
α
θ
z
w
N
M
Topic Model



Features are discourses

music song sing singing sang play played songs playing heard tune…

snow cold rain wind ice storm weather winter warm night air…

computer machine data system work program new information machines human computers…

LDA: Blei et al. NeurIPS 2002

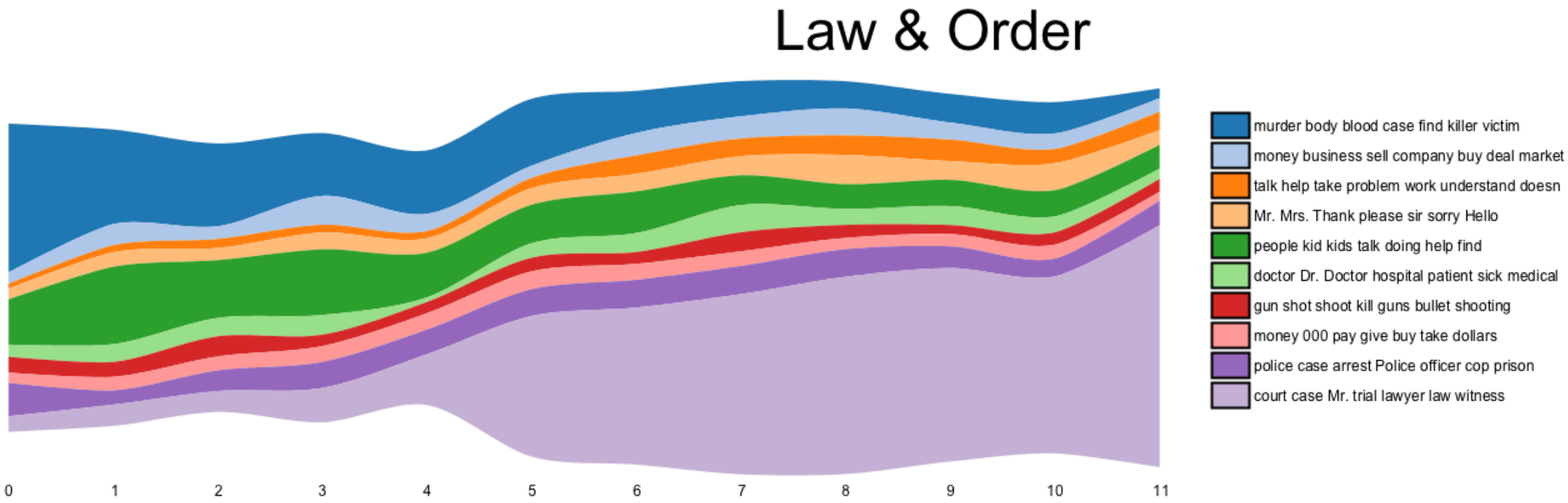
Topics map down to individual words

The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

Sea/Ocean: sea water boat island beach ship ocean …

Magic: magic spell witch power demon wizard magician …

Explore trends across “novel” time

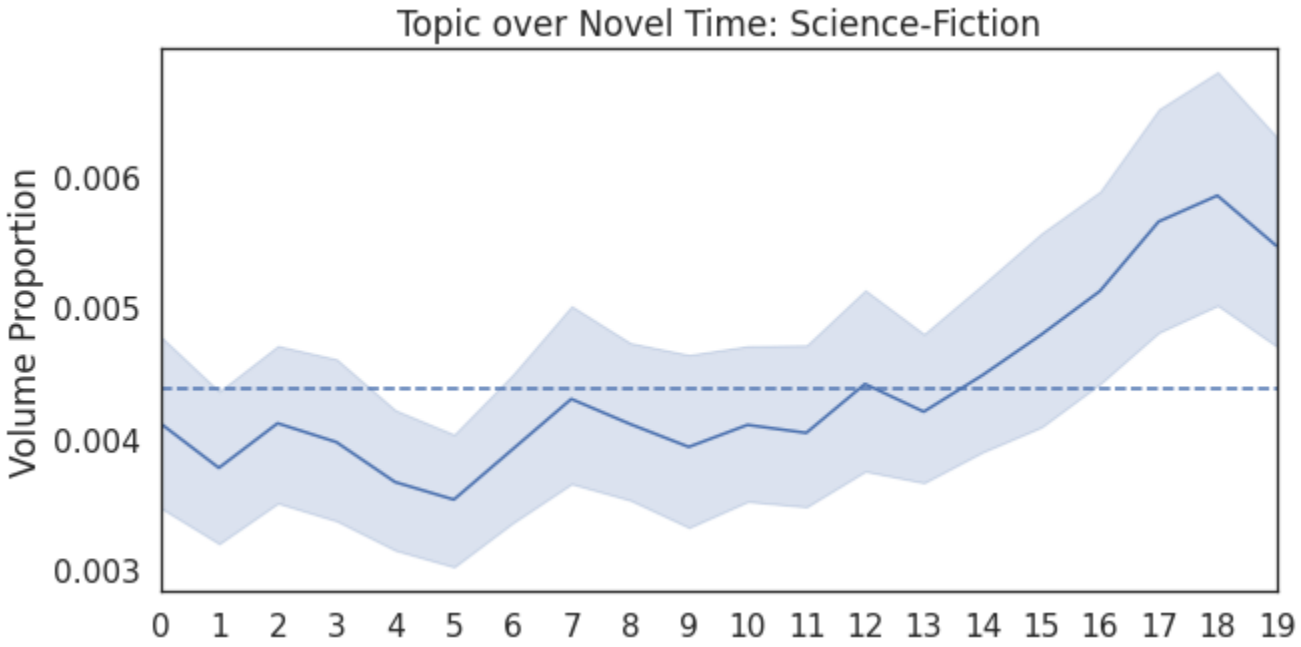


Schmidt Blog 12/2014

Science Fiction has spacecraft

ship ships fleet space planet our system vessel battle captain attack crew star new enemy craft aboard weapons those

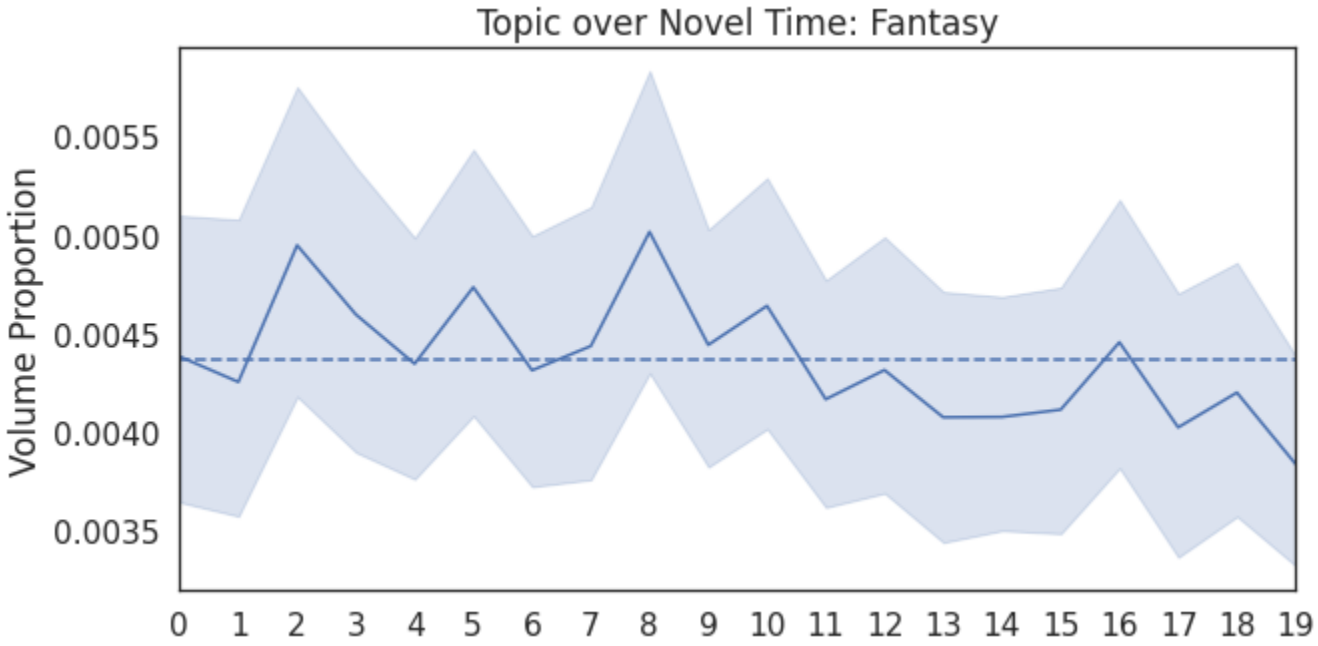
Top Novels	% of Vol.
Darksaber by Kevin J. Anderson	7.94%
Lyon's Pride by Anne McCaffrey	6.42%
Space Viking by H. Beam Pipe	5.99%
Outbound Flight by Timothy Zahn	5.13%



Fantasy has horses

horse horses rode ride saddle riding road behind reins men rider mare mount mounted wagon stable riders led hooves

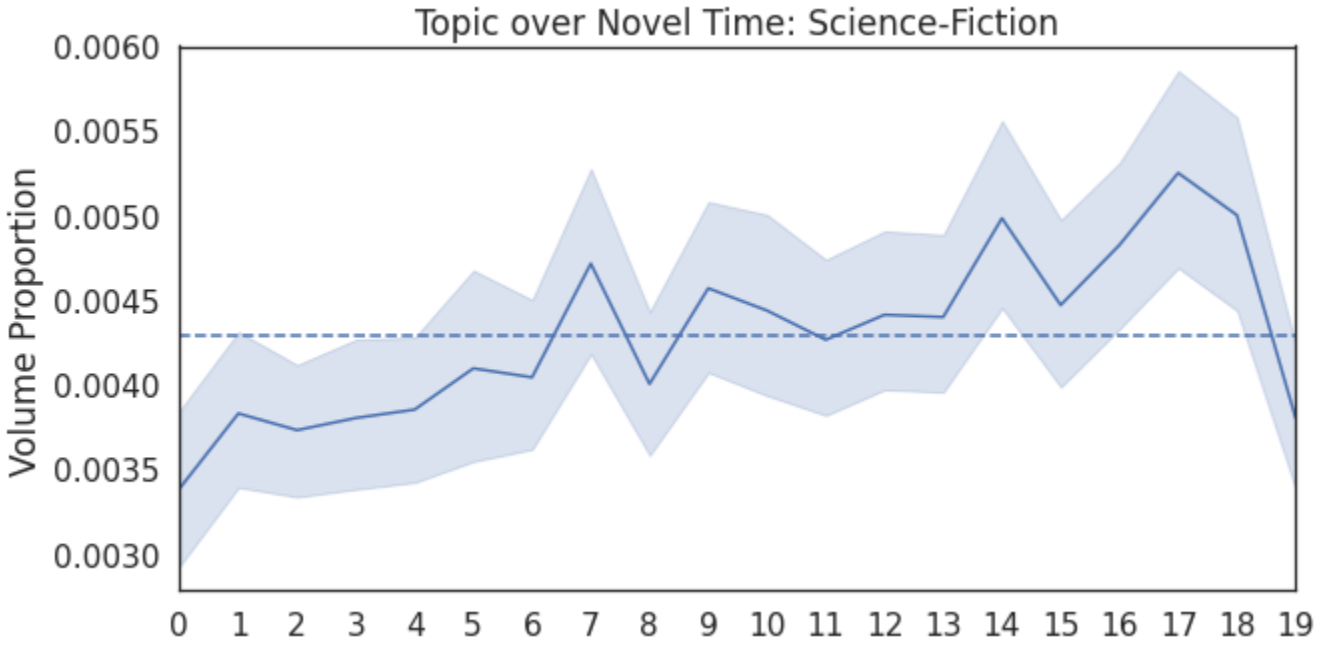
Top Novels	% of Vol.
Hawkmistress! by Marion Zimmer Bradley	4.32%
The Grey Horse by R. A. MacAvoy	4.00%
The Horse and His Boy by C. S. Lewis	3.84%
Watchtower by Elizabeth A. Lynn	3.14%



Science Fiction has radios

message radio signal contact send news station sent messages ship our report voice call communication has information

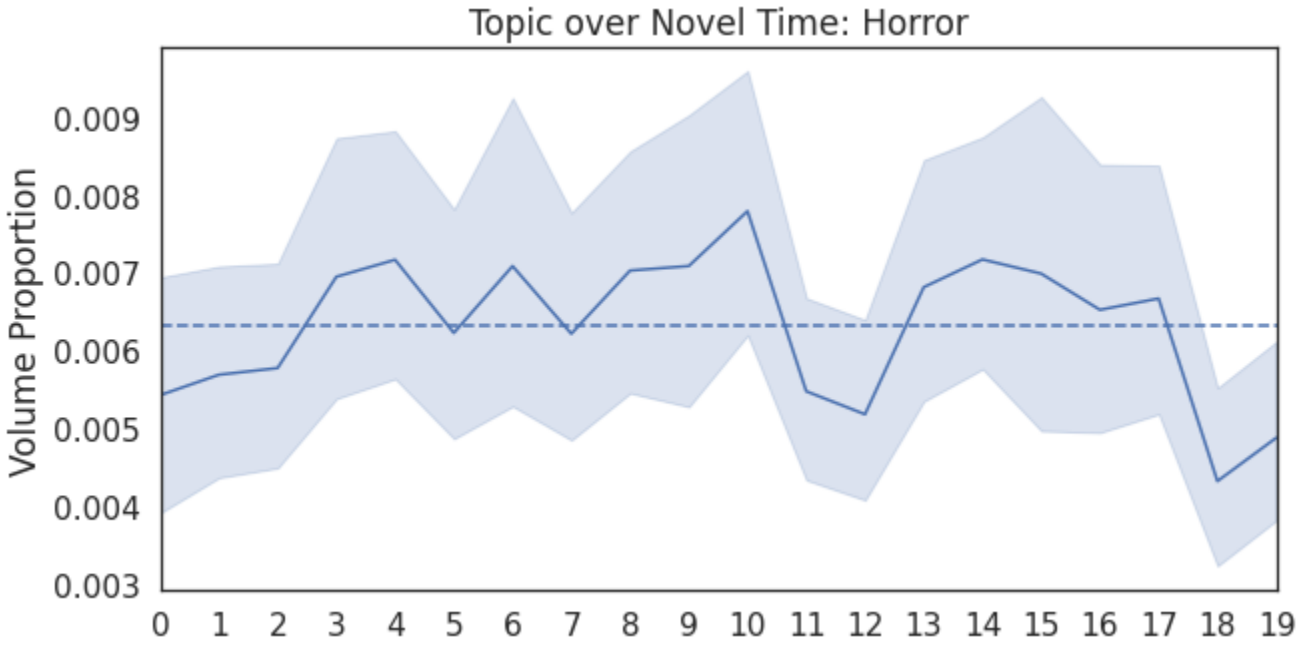
Top Novels	% of Vol.
The Pride of Chanur by C. J. Cherryh	2.80%
Explorer by C. J. Cherryh	2.64%
The Listeners by James E. Gunn	2.42%
The Andromeda Strain by Michael Crichton	2.29%



Horror has telephones

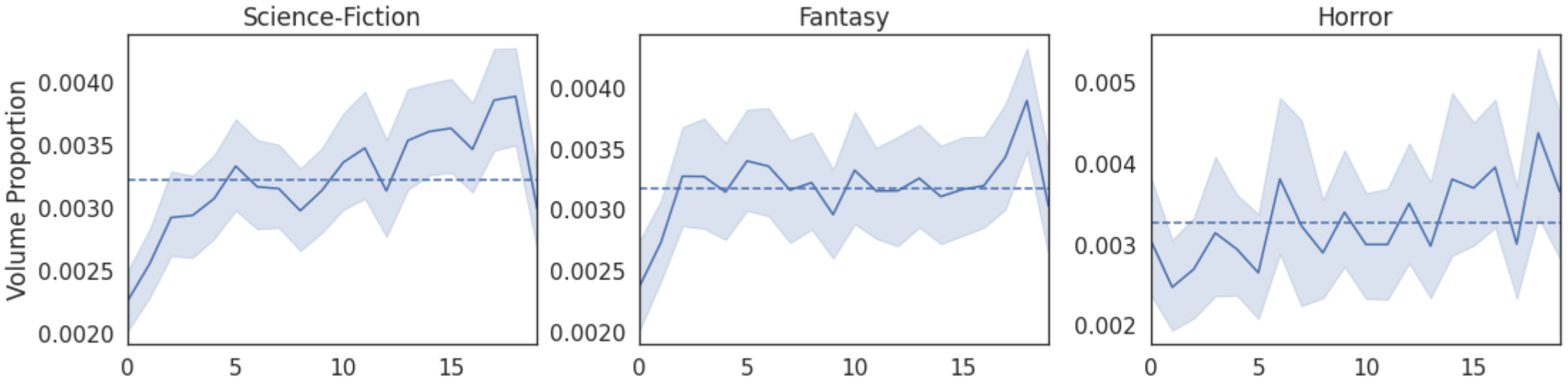
phone call number voice telephone called got rang hello yes line tell receiver calling hung want talk told ring picked calls

Top Novels	% of Vol.
Cold Heaven by Brian Moore	2.80%
Passage by Connie Willis	2.64%
The Dark Half by Stephen King	2.42%
Death is a Lonely Business by Ray Bradbury	2.29%



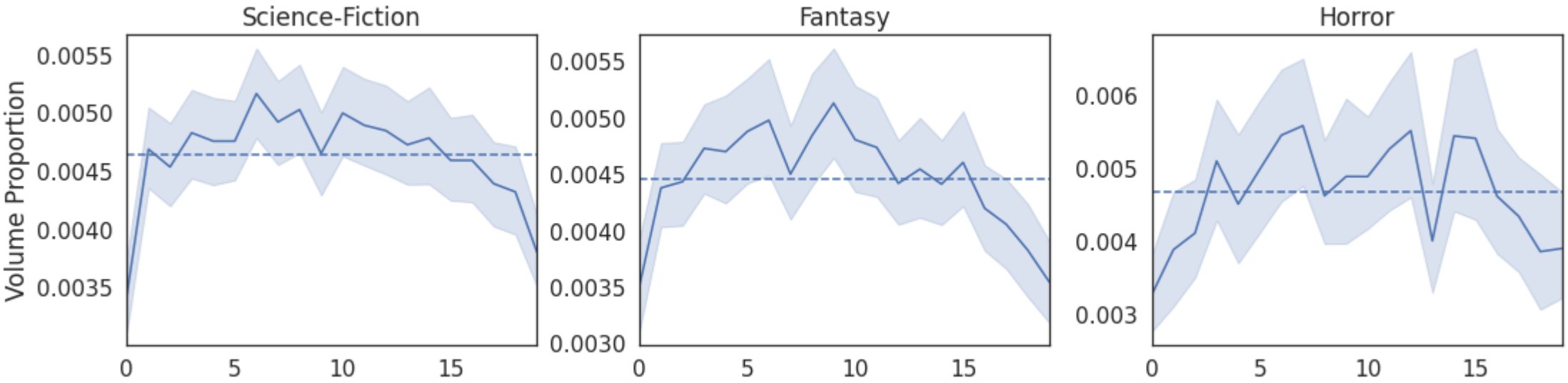
Pain is universal

pain hurt leg arm felt left broken body feel help tried side bad wound hand face injured himself badly feet while move legs better



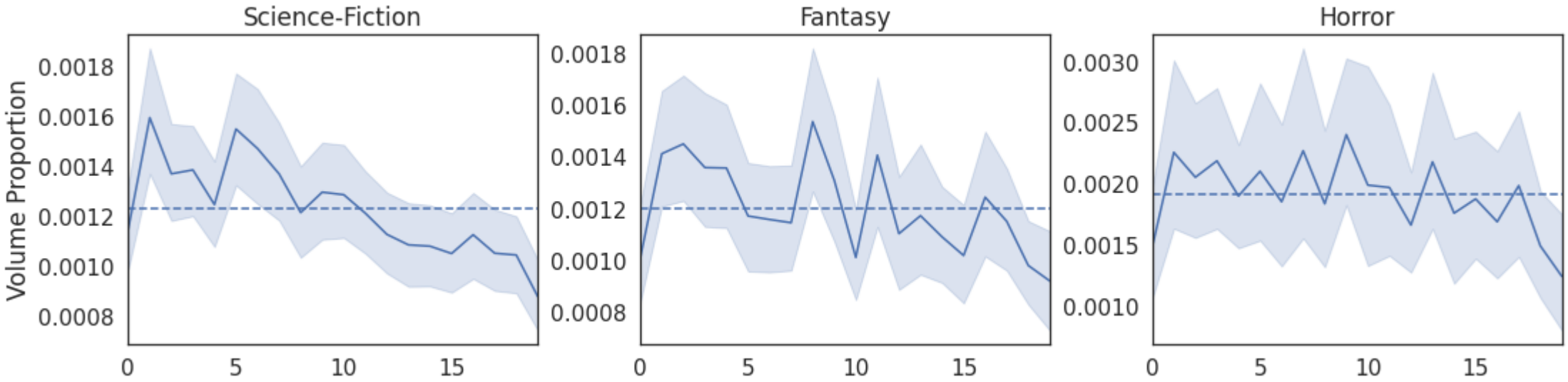
...so is information seeking/exchange

asked answer question why questions ask tell answered yes told asking answers wanted say knew want because many should mind



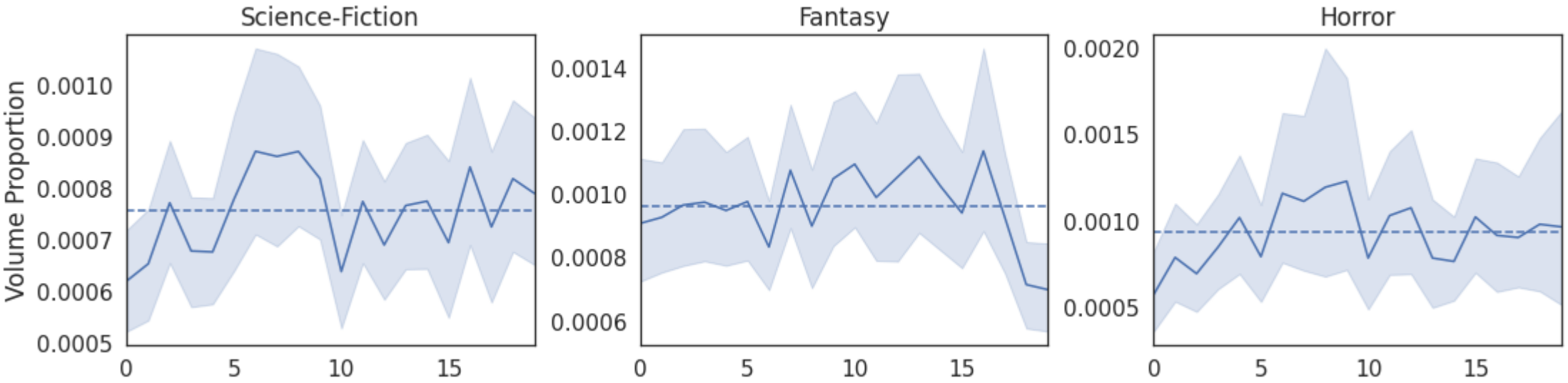
Smoking: Similar but different use

cigarette smoke pipe lit took smoking cigar cigarettes tobacco pack match smoked pocket sat lighter another light mouth blew hand



Creepy-Crawlies: Similar but different use

spider insects flies bees nest ants insect web spiders bugs snakes tiny swarm black buzzing bee fly creatures hive things small beetle

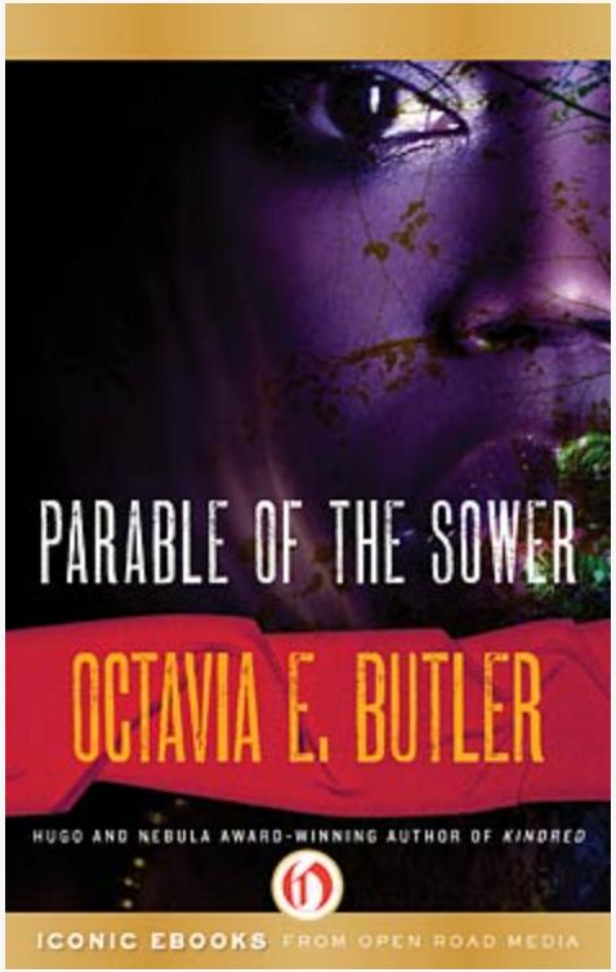


Combine: Libraries + Fan Databases

HATHI TRUST

Combine: Libraries + Fan Databases

HATHI TRUST



Added By: Administrator
Last Updated: Administrator

Parable of the Sower

Author:	Octavia E. Butler
Publisher:	Open Road Integrated Media, 2012 Four Walls Eight Windows, 1993
⊞ Series:	The Parable Series: Book 1
Book Type:	Novel
Genre:	Science-Fiction
Sub-Genre Tags:	Apocalyptic/Post-Apocalyptic Near-Future Dystopia

If you liked **Parable of the Sower** you might like these books.

Awards:	• 1994 Nebula Nominated • 1995 Locus SF Nominated
Lists:	• The Classics of Science Fiction • Science Fiction: The 101 Best Novels 1985-2010 • WWEnd Top Listed Books of All-Time • The Defining Science Fiction Books of the 1990s • 200 Significant SF Books by Women, 1984-2001 • Science Fiction by Women Writers

Building a curated list with 

List: 18,809 works by 3,620 authors published from 1900–2010

Book Type	# Works
Novel	13,585
Collection	1,713
Anthology	1,428
Omnibus	426
Novella	1003
Novelette	654

Genre	# Works
Science-Fiction	10,092
Fantasy	6,347
Horror	1,557
Science-Fiction / Fantasy	419
Fantasy / Horror	301
Science-Fiction / Horror	54
S / F / H	39

Searching the HathiTrust Catalog

Found: 3,241 works & 5,160 volumes

Book Type	# Works	# Volumes
Novel	2,374	3,862
Collection	436	674
Anthology	376	517
Omnibus	30	47
Novella	23	51
Novelette	2	9

Genres have similar match rates

Genre	% Matched	# Matched
Science-Fiction	18.6%	2,801
Fantasy	15.1%	1,755
Horror	17.3%	404

…but subgenre coverage varies widely

Highest

Subgenre	Prop. Matched	% Matched
Theological	79/140	**56.4%**
Soft SF	57/122	46.7%
Dying Earth	46/100	46.0%
Human Development	130/296	43.9%
Magical Realism	56/129	43.4%

Lowest

Subgenre	Prop. Matched	% Matched
Urban Fantasy	20/197	**10.2%**
Sword & Sorcery	24/192	12.5%
Space Opera	125/917	13.6%
Vampires	29/178	16.3%
Military SF	54/282	19.1%

HathiTrust contains many compilations

Anthologies and collections have highest match rates

Book Type	% Matched	# Matched
Anthology	**26.3%**	376
Collection	**25.5%**	436
Novel	17.5%	2,374

Worlds Without End has more limited coverage of compilations

250+ additional anthologies found!

Top authors are fairly prolific

Author	Prop. Matched	% Matched
Robert Silverberg	47 / 138	34.1%
Robert A. Heinlein	40 / 49	81.6%
Isaac Asimov	38 / 67	56.7%
Michael Moorcock	38 / 95	40.0%
Andre Norton	38 / 112	33.9%

...but some prolific authors have few matches

Missing

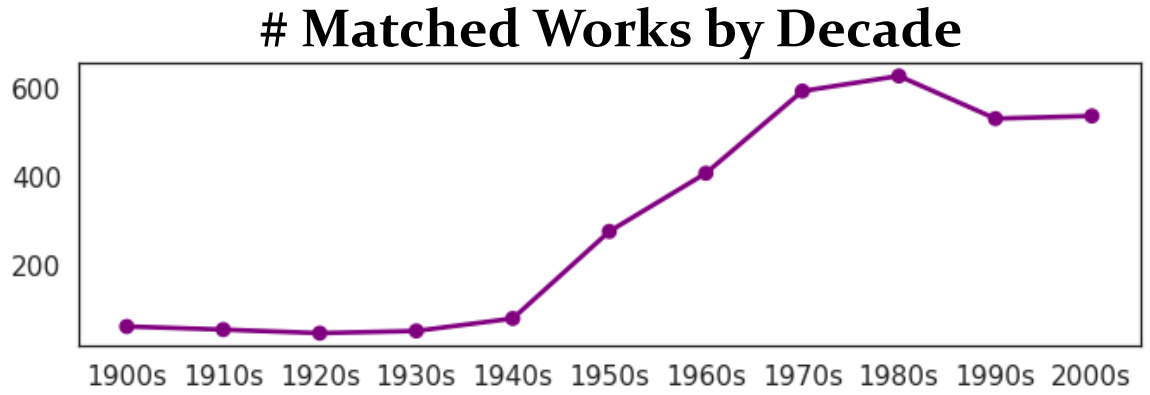
Author	# Works
James Axler	94
Brian Stableford	62
L. E. Modesitt, Jr.	61
Jack L. Chalker	54
E. C. Tubb	52

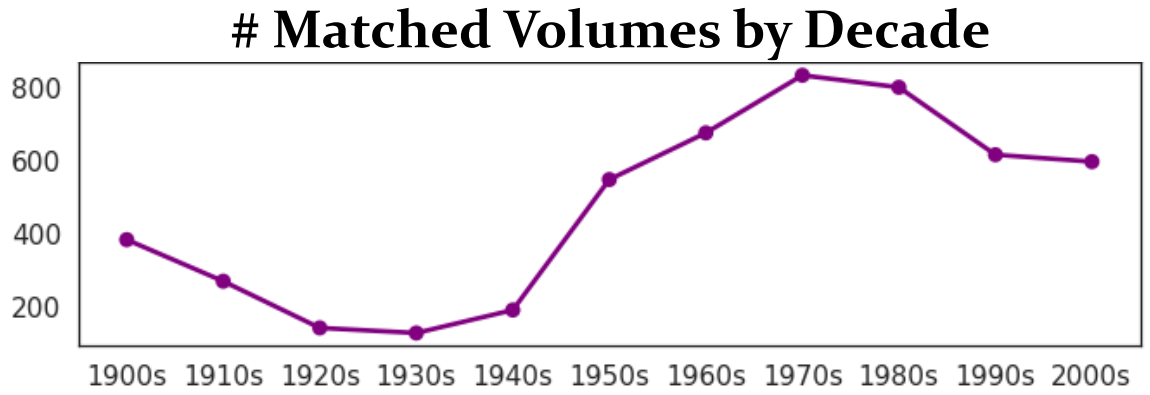
Few Matches

Author	Coverage
Fred Saberhagen	2 / 55
Glen Cook	3 / 54
Chelsea Quinn Yarbro	3 / 49
R. A. Salvatore	4 / 50
Mercedes Lackey	5 / 55

Public domain works are overrepresented

Novel	Year	# Vols.
Just So Stories by Rudyard Kipling	1902	31
The Wind in the Willows by Kenneth Grahame	1908	25
Zuleika Dobson by Max Beerbohm	1911	25
Before Adam by Jack London	1906	21
Puck of Pook's Hill by Rudyard Kipling	1906	21

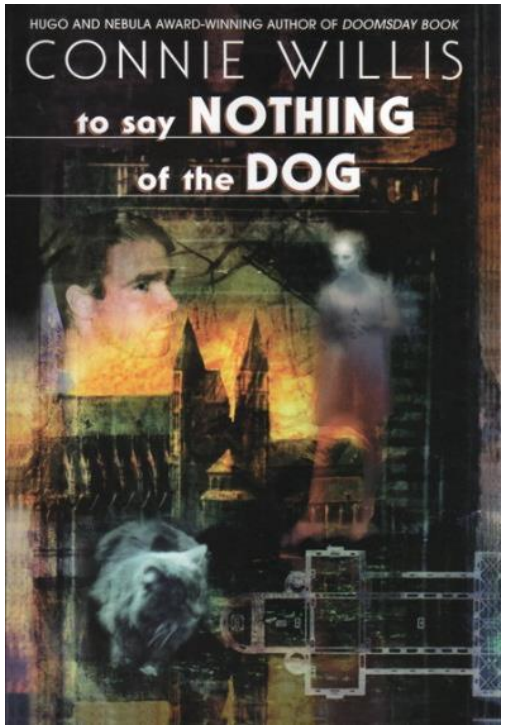


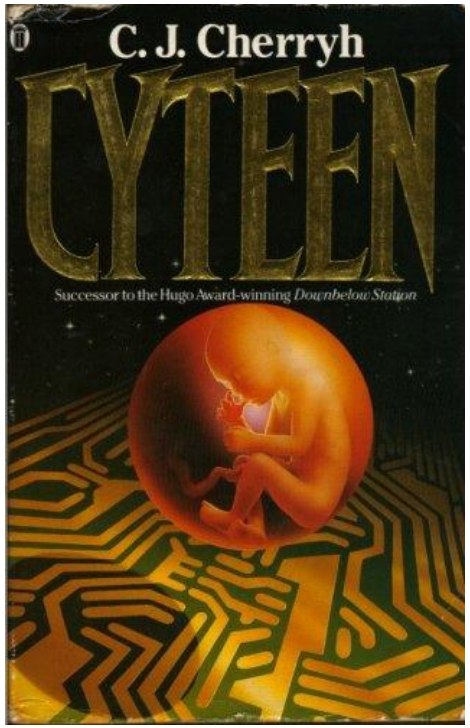


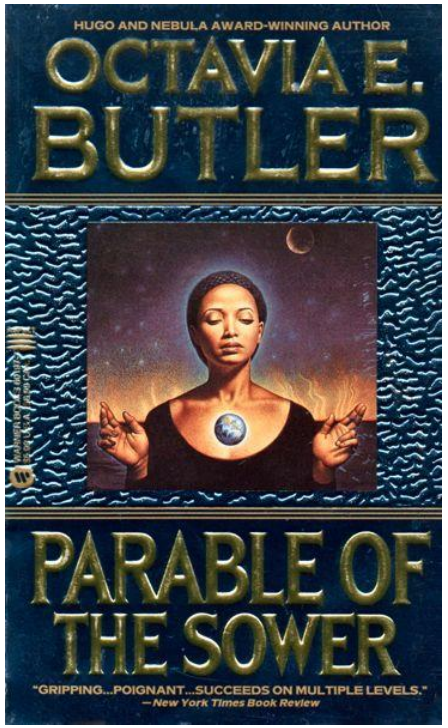
Award-winning works have better coverage

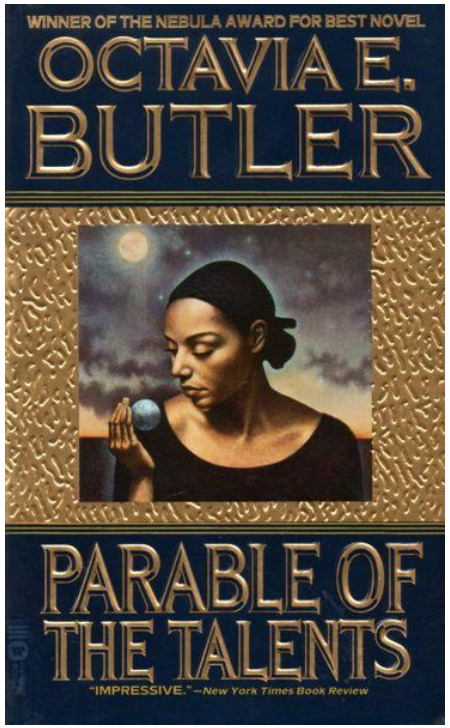
Award	% of Winners	% of Nominees
Hugo	63.9%	51.1%
Nebula	61.7%	46.6%
Bram Stoker	46.2%	27.7%
Locus Science Fiction	57.1%	50.0%
Locus Fantasy	45.5%	35.2%

...but highly regarded works are still absent

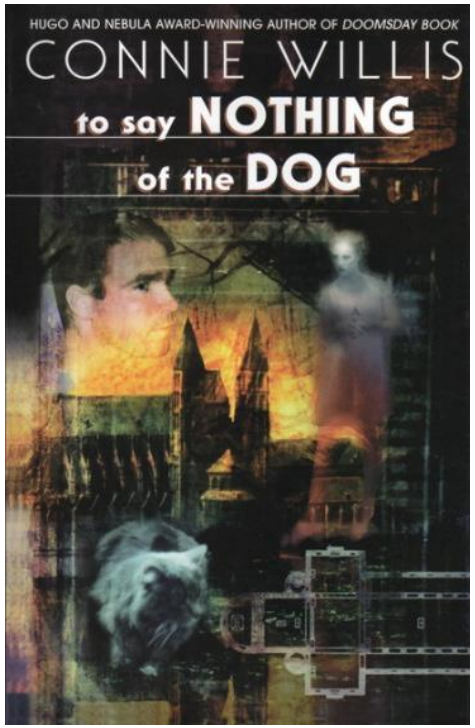






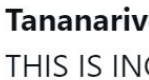


...but highly regarded works are still absent



Tananarive (Team Pfizer) Due @TananariveDue · Sep 2, 2020

THIS IS INCREDIBLE. Octavia Butler's agent has posted that she is finally a NY Times bestselling author, one of Octavia's goals! FANTASTIC!!!



Merrilee Heifetz @MerrileeHeifetz · Sep 2, 2020

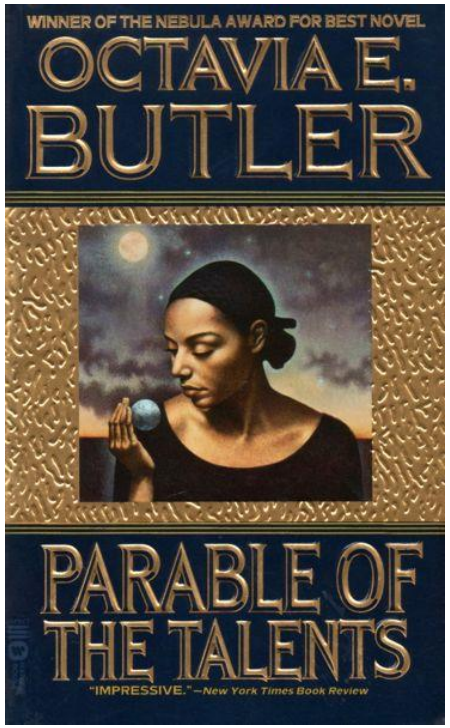
Octavia E. Butler, who died in 2006, is a NYT Bestselling author. This was one of her life goals. Thank you all for making it happen!

12 **NORMAL PEOPLE,** by Sally Rooney. (Hogarth) The connection between a high school star athlete and a loner ebbs and flows when they go to Trinity College in Dublin.

13 **THE NICKEL BOYS,** by Colson Whitehead. (Anchor) Two boys respond to horrors at a Jim Crow-era reform school in ways that impact them decades later.

14 **PARABLE OF THE SOWER,** by Octavia E. Butler. (Grand Central) Fifteen-year-old Lauren Olamina fights to have her voice heard in her California community beset by climate change and economic crises.

15 **THE OVERSTORY,** by Richard Powers. (Norton) Winner of the 2019 Pulitzer Prize for fiction. Nine people drawn to trees for different reasons fight for the last of the remaining acres of virgin forest.



WWEnd lists highlight disparities

WWEnd List	% Matched
The Defining Science Fiction Books of the 1950s	77.6%
The Defining Science Fiction Books of the 1960s	71.4%
The Defining Science Fiction Books of the 1970s	74.5%
The Defining Science Fiction Books of the 1980s	60%
The Defining Science Fiction Books of the 1990s	27.2%

WWEnd lists highlight disparities

WWEnd List	% Matched
Science Fiction Masterworks	68.4%
Fantasy Masterworks	40.6%
Science Fiction Mistressworks	39.4%

Searching catalogue records is hard!

85% of matched volumes identified by automated catalogue search

Complement with content-based methods

Sim.	HTID	Title / Author	Year
1	nyp.33433112045251	The chessmen of Mars / by Edgar Rice Burroughs ... ; illustrated by J. Allen St. John	1922
0.988	osu.32435017883182	The chessmen of Mars / by Edgar Rice Burroughs ... ; illustrated by J. Allen St. John.	1922
0.802	osu.32435017174004	Thuvia, maid of Mars / by Edgar Rice Burroughs, illustrated by J. Allen St. John.	1920

Sim.	HTID	Title / Author	Year
1	mdp.39015013315810	The Hugo winners, edited by Isaac Asimov.	9999
0.963	pst.000012384754	The Science fiction hall of fame.	9999
0.962	mdp.39015000656127	Dangerous visions; 33 original stories. Illus. by Leo and Diane Dillon.	1967

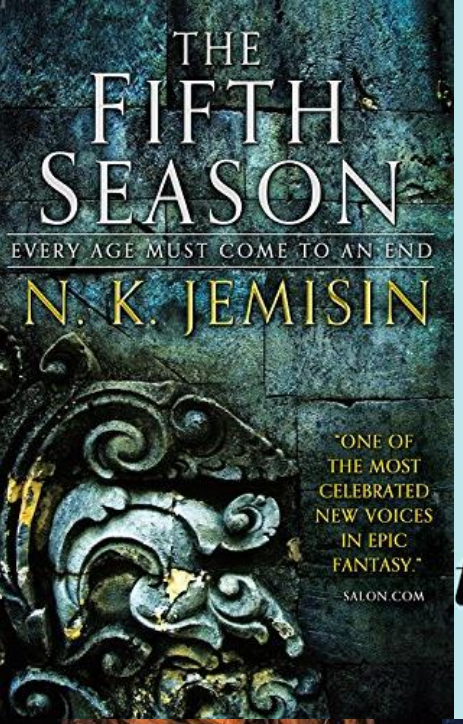
HathiTrust is a powerful tool for studying speculative fiction!

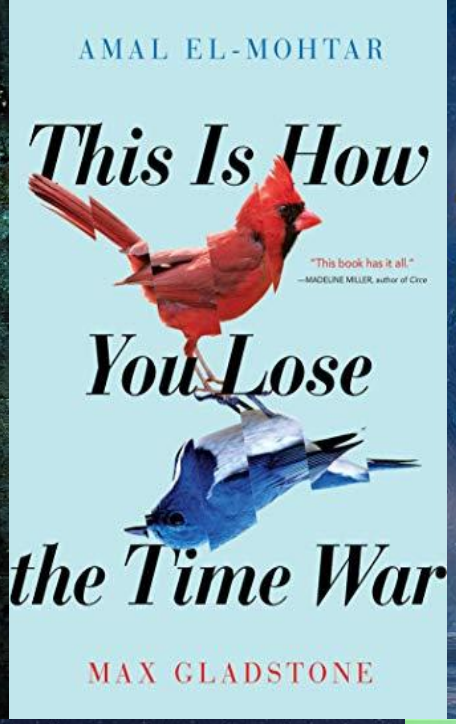
HathiTrust Recommended Workset:

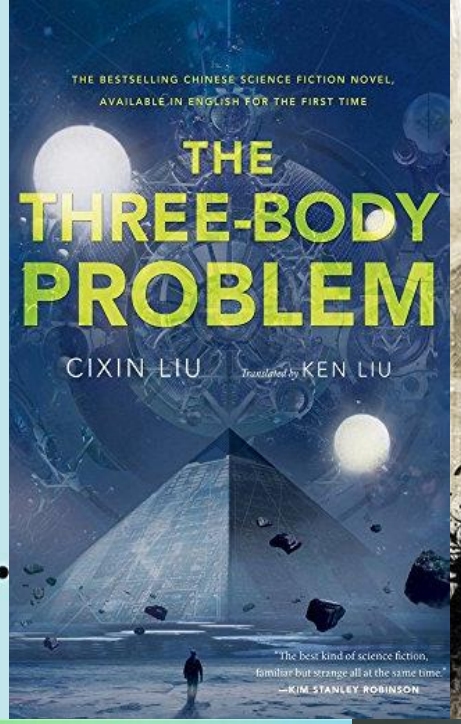
"20th Century English-Language Speculative Fiction"

See: analytics.hathitrust.org/staticrecommendedworksets

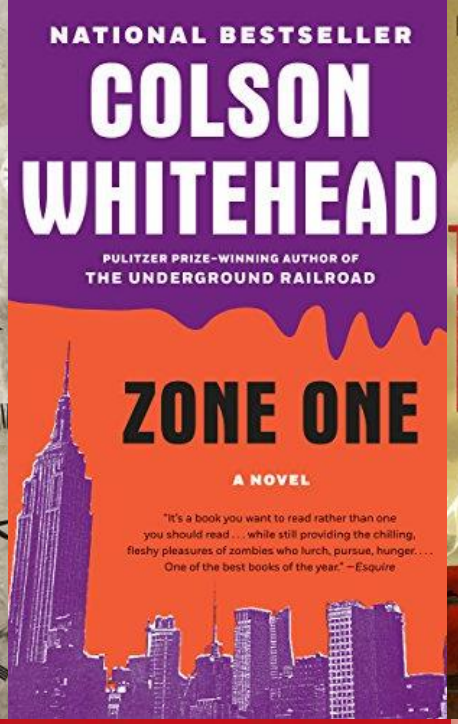
Github: github.com/laurejt/sf-in-hathitrust

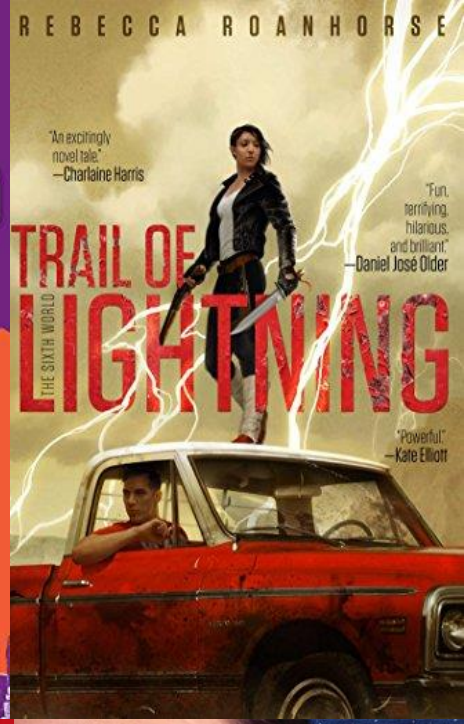

THE FIFTH SEASON
EVERY AGE MUST COME TO AN END
N. K. JEMISIN
"ONE OF THE MOST CELEBRATED NEW VOICES IN EPIC FANTASY."
SALON.COM

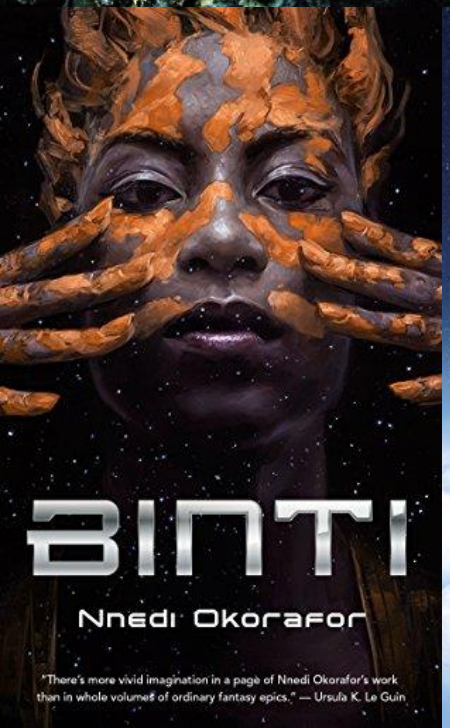

AMAL EL-MOHTAR
This Is How You Lose the Time War
"This book has it all."
—MADELINE MILLER, author of Circe
MAX GLADSTONE

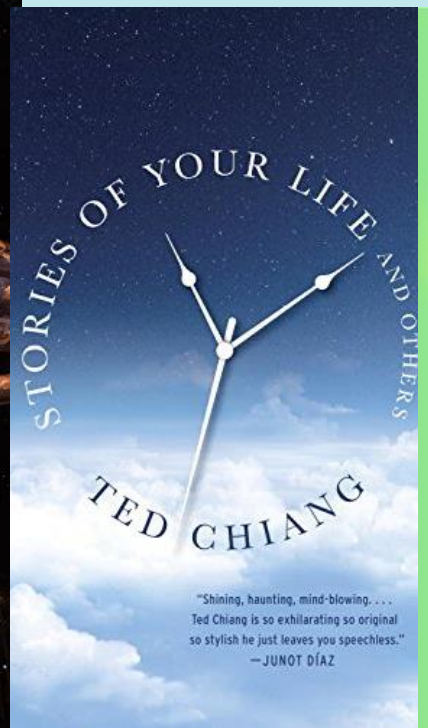

THE BESTSELLING CHINESE SCIENCE FICTION NOVEL, AVAILABLE IN ENGLISH FOR THE FIRST TIME
THE THREE-BODY PROBLEM
CIXIN LIU
Translated by KEN LIU
"The best kind of science fiction, familiar but strange all at the same time."
—KIM STANLEY ROBINSON

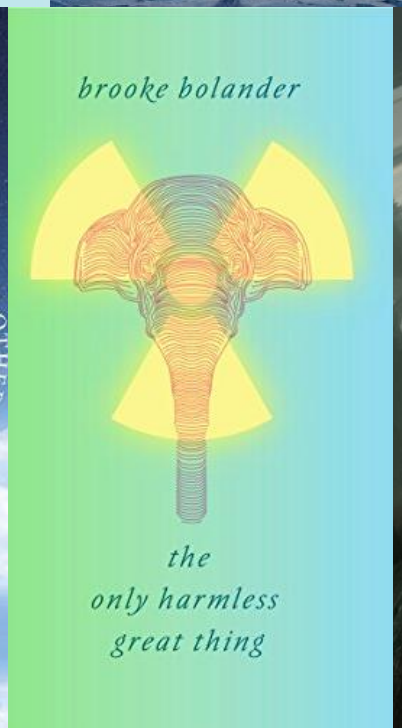

CONNIE WILLIS
Doomsday Book
WINNER OF THE HUGO & NEBULA AWARDS

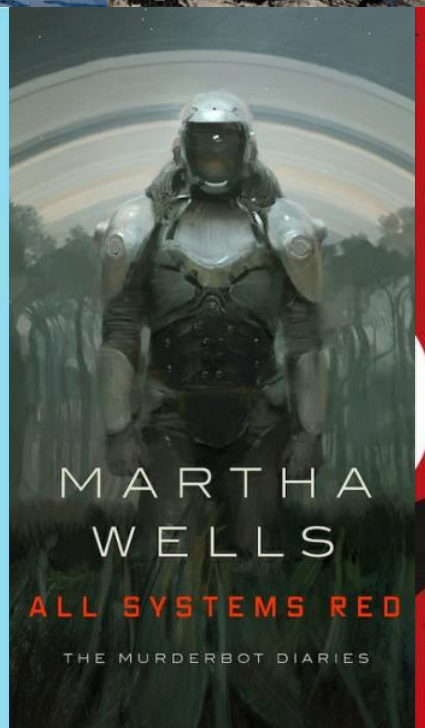

NATIONAL BESTSELLER
COLSON WHITEHEAD
PULITZER PRIZE-WINNING AUTHOR OF THE UNDERGROUND RAILROAD
ZONE ONE
A NOVEL
"It's a book you want to read rather than one you should read . . . while still providing the chilling, fleshy pleasures of zombies who lurch, pursue, hunger. . . . One of the best books of the year." —Esquire

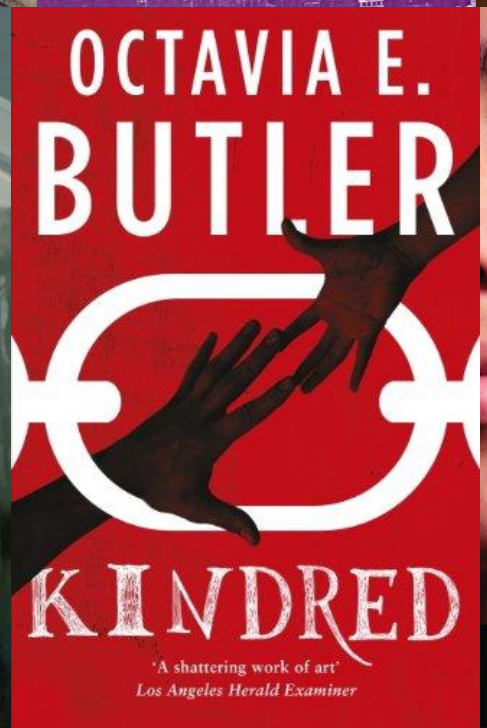

REBECCA ROANHORSE
"An exciting novel tale."
—Charlaine Harris
TRAIL OF LIGHTNING
THE SIXTH WORLD
"Fun, terrifying, hilarious, and brilliant."
—Daniel José Older
"Powerful."
—Kate Elliott

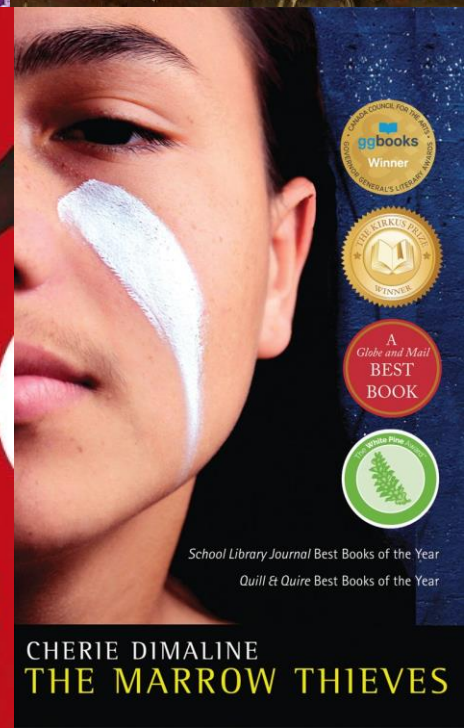

BINTI
Nnedi Okorafor
"There's more vivid imagination in a page of Nnedi Okorafor's work than in whole volumes of ordinary fantasy epics." — Ursula K. Le Guin


STORIES OF YOUR LIFE AND OTHERS
TED CHIANG
"Shining, haunting, mind-blowing. . . . Ted Chiang is so exhilarating so original so stylish he just leaves you speechless."
—JUNOT DÍAZ


brooke bolander
the only harmless great thing


MARTHA WELLS
ALL SYSTEMS RED
THE MURDERBOT DIARIES


OCTAVIA E. BUTLER
KINDRED
'A shattering work of art'
Los Angeles Herald Examiner


ggbooks Winner
A Globe and Mail BEST BOOK
School Library Journal Best Books of the Year
Quill & Quire Best Books of the Year
CHERIE DIMALINE
THE MARROW THIEVES