Best selling Books: Data Analysis

Weekly Update
https://github.com/sruthipsuresh/best-sellers-analysis

Weekly Summary

- Wanted to create a “pipeline” of analysis scripts which can then be used for larger datasets going back a year as discussed.
- Worked with 2 datasets sourced from the latest NYTBL (non-fiction) and Amazon non-fiction (paid) best sellers - learned how to use Google Books API and work with json as intermediate processing step for Amazon due to efficiency issues.
- Analyzed both titles and descriptions as a “proof of concept”.

  Polarity/subjectivity (textblob), EDA with Wordcloud, Topic Modeling, Clustering

Data Collection and Cleaning

Amazon and Combined NYT

BestSellers2021Scraping.ipynb > gBooksfinal.ipynb > gbooksprocessing.ipynb

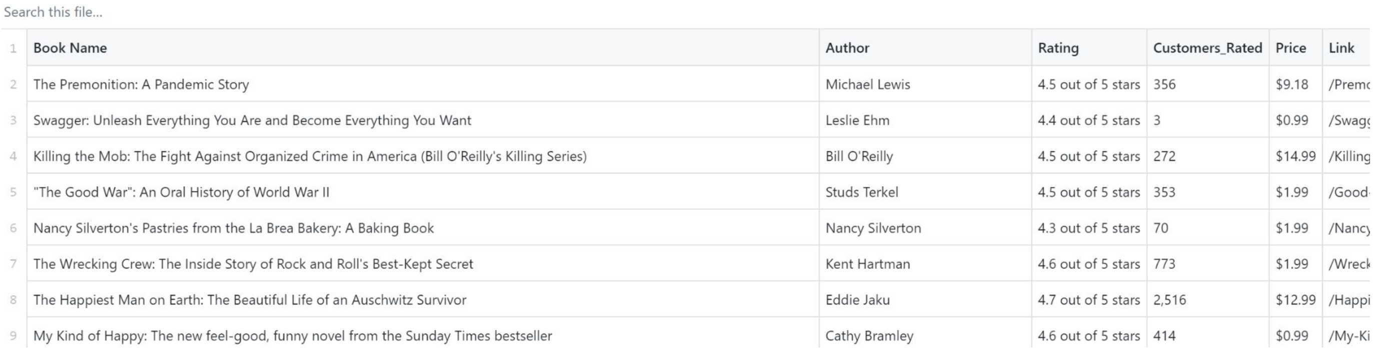
NYT only

NYT_Bestseller_Filter.ipynb > nytcleaning.ipynb

Data Collection (Nonfiction Books):

AMAZON

Collected all top-selling books in Amazon Best Sellers: Nonfiction. (https://www.amazon.com/Best-Sellers-Kindle-Store-Nonfiction/zgbs/digital-text/157325011/ref=zg_bs)



Search this file...

Book Name	Author	Rating	Customers_Rated	Price	Link
The Premonition: A Pandemic Story	Michael Lewis	4.5 out of 5 stars	356	$9.18	/Prem
Swagger: Unleash Everything You Are and Become Everything You Want	Leslie Ehm	4.4 out of 5 stars	3	$0.99	/Swag
Killing the Mob: The Fight Against Organized Crime in America (Bill O'Reilly's Killing Series)	Bill O'Reilly	4.5 out of 5 stars	272	$14.99	/Killing
"The Good War": An Oral History of World War II	Studs Terkel	4.5 out of 5 stars	353	$1.99	/Good
Nancy Silverton's Pastries from the La Brea Bakery: A Baking Book	Nancy Silverton	4.3 out of 5 stars	70	$1.99	/Nanc
The Wrecking Crew: The Inside Story of Rock and Roll's Best-Kept Secret	Kent Hartman	4.6 out of 5 stars	773	$1.99	/Wrec
The Happiest Man on Earth: The Beautiful Life of an Auschwitz Survivor	Eddie Jaku	4.7 out of 5 stars	2,516	$12.99	/Happi
My Kind of Happy: The new feel-good, funny novel from the Sunday Times bestseller	Cathy Bramley	4.6 out of 5 stars	414	$0.99	/My-Ki

NYBT

Books in the combined book and

e-book list (latest).

GBooks API Issues

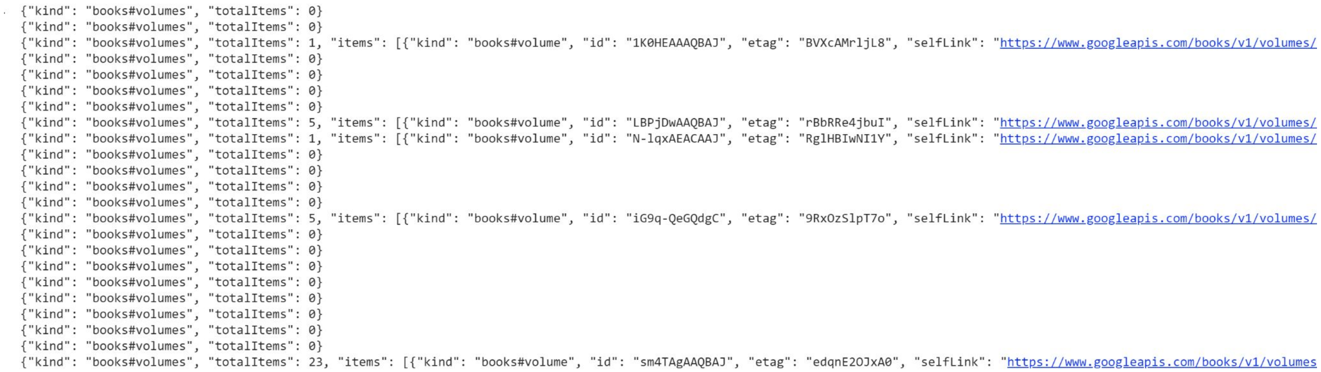
AMAZON BOOKS

29/50 Amazon books had an exact

match in Google Books API.

NYT BOOKS

7 books had exact match.

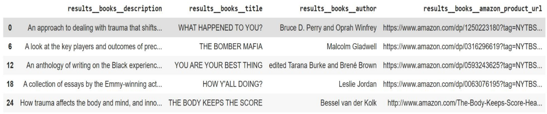


```
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 1, "items": [{"kind": "books#volume", "id": "1K0HEAAAQBAJ", "etag": "BVXcAMrljL8", "selfLink": "https://www.googleapis.com/books/v1/volumes/
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 5, "items": [{"kind": "books#volume", "id": "LBPjDwAAQBAJ", "etag": "rBbRRe4jbuI", "selfLink": "https://www.googleapis.com/books/v1/volumes/
{"kind": "books#volumes", "totalItems": 1, "items": [{"kind": "books#volume", "id": "N-lqxAEACAAJ", "etag": "RglHBIwNI1Y", "selfLink": "https://www.googleapis.com/books/v1/volumes/
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 5, "items": [{"kind": "books#volume", "id": "iG9q-QeGQdgC", "etag": "9RxOzSlpT7o", "selfLink": "https://www.googleapis.com/books/v1/volumes/
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 0}
{"kind": "books#volumes", "totalItems": 23, "items": [{"kind": "books#volume", "id": "sm4TAgAAQBAJ", "etag": "edqnE2OJxA0", "selfLink": "https://www.googleapis.com/books/v1/volumes
```

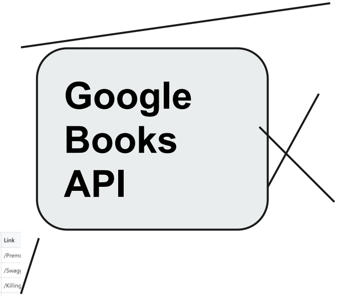
Chose this as a medium as not all books in Amazon list in NYT database.

Datasets (Nonfiction Books):

NYT BOOKS



	results__books__description	results__books__title	results__books__author	results__books__amazon_product_url
0	An approach to dealing with trauma that shifts...	WHAT HAPPENED TO YOU?	Bruce D. Perry and Oprah Winfrey	https://www.amazon.com/dp/1250223180?tag=NYTBS...
6	A look at the key players and outcomes of prec...	THE BOMBER MAFIA	Malcolm Gladwell	https://www.amazon.com/dp/0316296619?tag=NYTBS...
12	An anthology of writing on the Black experienc...	YOU ARE YOUR BEST THING	edited Tarana Burke and Brené Brown	https://www.amazon.com/dp/0593243625?tag=NYTBS...
18	A collection of essays by the Emmy-winning act...	HOW Y'ALL DOING?	Leslie Jordan	https://www.amazon.com/dp/0063076195?tag=NYTBS...
24	How trauma affects the body and mind, and inno...	THE BODY KEEPS THE SCORE	Bessel van der Kolk	http://www.amazon.com/The-Body-Keeps-Score-Hea...



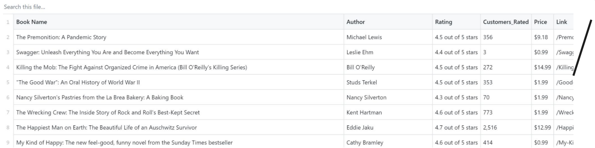
Google Books API

NYT ONLY (16 books from most recent list + NYT descriptions)

NYT w/ Google Books API (7 books from most recent list + Books descriptions)

AMAZON ONLY (29 books descriptions lifted from API)

AMAZON BOOKS



Search this file...

Book Name	Author	Rating	Customers_Rated	Price	Link
The Premonition: A Pandemic Story	Michael Lewis	4.5 out of 5 stars	356	$9.18	/Premo
Swagger: Unleash Everything You Are and Become Everything You Want	Leslie Ehm	4.4 out of 5 stars	3	$0.99	/Swag
Killing the Mob: The Fight Against Organized Crime in America (Bill O'Reilly's Killing Series)	Bill O'Reilly	4.5 out of 5 stars	272	$14.99	/Killing
"The Good War": An Oral History of World War II	Studs Terkel	4.5 out of 5 stars	353	$1.99	/Good
Nancy Silverton's Pastries from the La Brea Bakery: A Baking Book	Nancy Silverton	4.3 out of 5 stars	70	$1.99	/Nancy
The Wrecking Crew: The Inside Story of Rock and Roll's Best-Kept Secret	Kent Hartman	4.6 out of 5 stars	773	$1.99	/Wreck
The Happiest Man on Earth: The Beautiful Life of an Auschwitz Survivor	Eddie Jaku	4.7 out of 5 stars	2,516	$12.99	/Happi
My Kind of Happy: The new feel-good, funny novel from the Sunday Times bestseller	Cathy Bramley	4.6 out of 5 stars	414	$0.99	/My-Ki

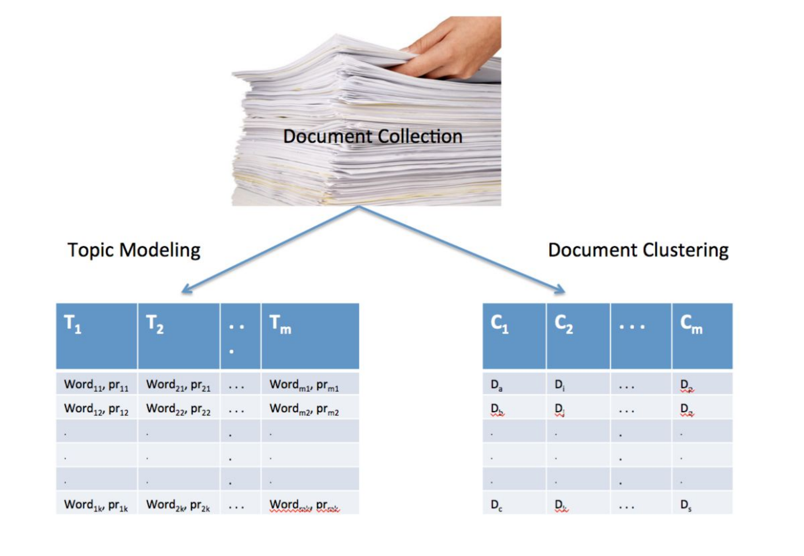
Chose this as a medium as not all books in Amazon list in NYT database. (Future: BOTH NYT and AMAZON will have descriptions from Google Books API)

Data Analysis

clustering.ipynb
topicmodeling.ipynb
wordcloudandtextblob.ipynb

https://iksinc.online/2016/05/16/topic-modeling-and-document-clustering-whats-the-difference/#:~:text=Irrespective%20of%20the%20approach%2C%20the,on%20some%20suitable%20similarity%20measure.

Topic Modeling v. Clustering



T_1	T_2	...	T_m
$Word_{11}$, pr_{11}	$Word_{21}$, pr_{21}	...	$Word_{m1}$, pr_{m1}
$Word_{12}$, pr_{12}	$Word_{22}$, pr_{22}	...	$Word_{m2}$, pr_{m2}
.	.	.	.
.	.	.	.
.	.	.	.
$Word_{1k}$, pr_{1k}	$Word_{2k}$, pr_{2k}	...	$Word_{mk}$, pr_{mk}

C_1	C_2	...	C_m
D_a	D_i	...	D_p
D_b	D_j	...	D_q
.	.	.	.
.	.	.	.
.	.	.	.
D_c	D_k	...	D_s

In this case,

Document Collection = Book Titles or Descriptions.

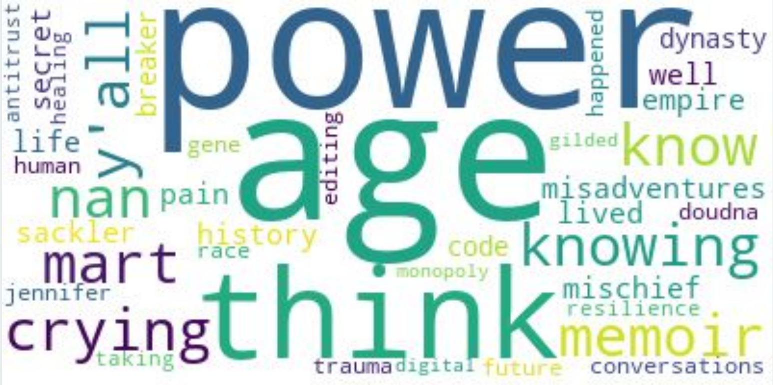
Topic Modeling: “ Each document is considered a mixture of topics and each word in a document is considered randomly drawn from document’s topics”

Document Clustering: Clusters of book titles/descriptions based on similarity measure.

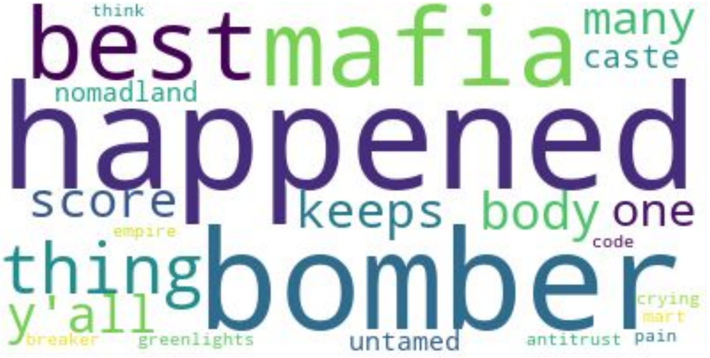
Title Text Analysis

Title Analysis: Exploratory Data Analysis using Wordcloud

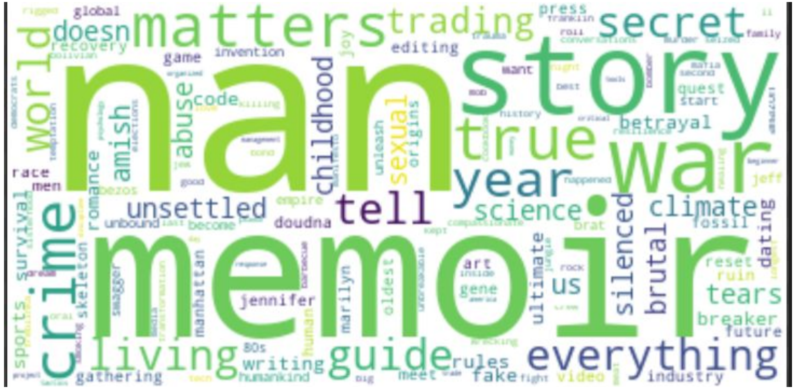
NYT + Gbooks



NYT ONLY

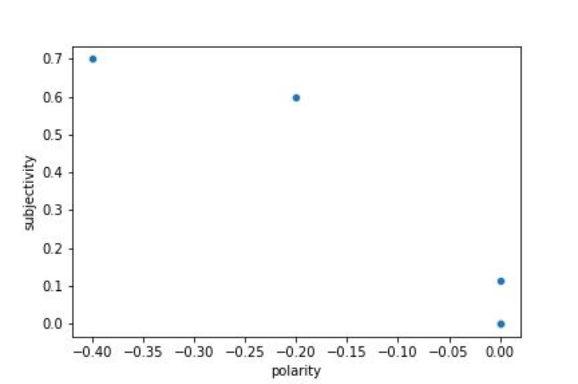


Amazon

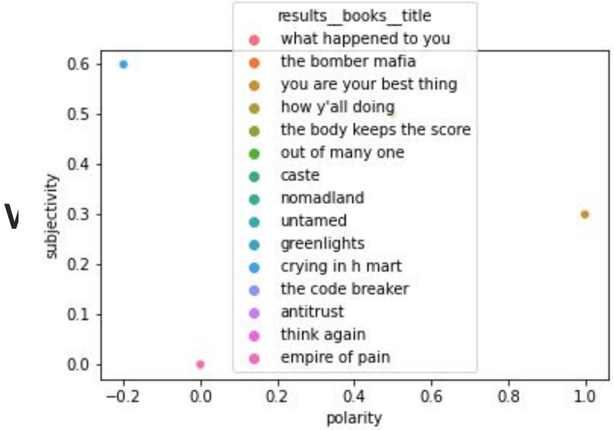


Title Analysis: Textblob Sentiment Analysis

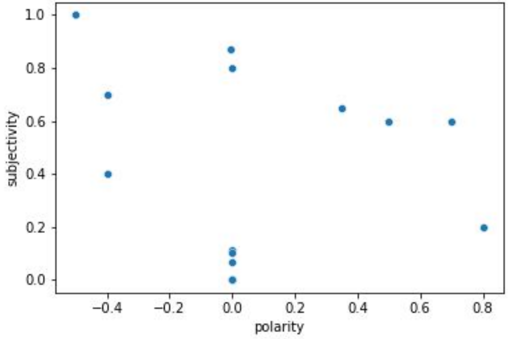
NYT + Gbooks



NYT ONLY



AMAZON



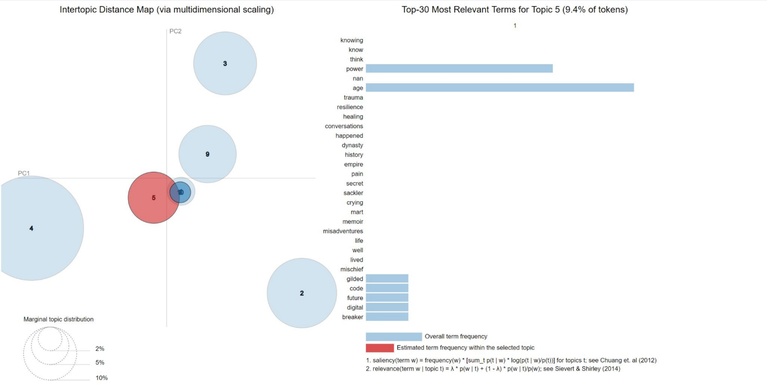
Title Analysis: Topic Modeling

Used:

https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0 (LDA topic modeling)

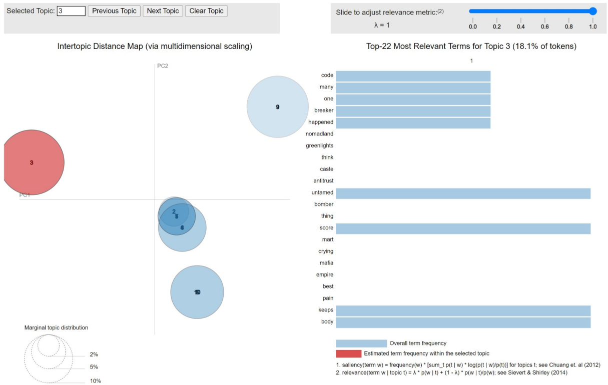
Title Analysis: LDA Topic Modeling (visualized via pyLDAvis)

NYT + Gbooks



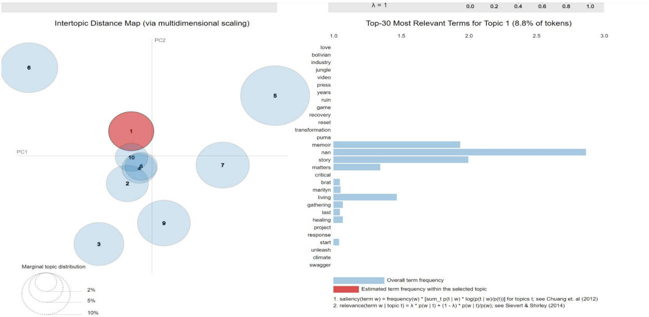
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

NYT ONLY



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

AMAZON



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

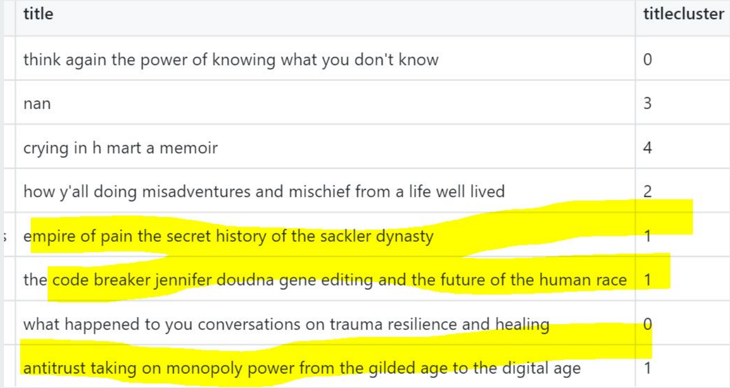
Title Analysis: Clustering

Used:

https://towardsdatascience.com/clustering-documents-with-python-97314ad6a78d

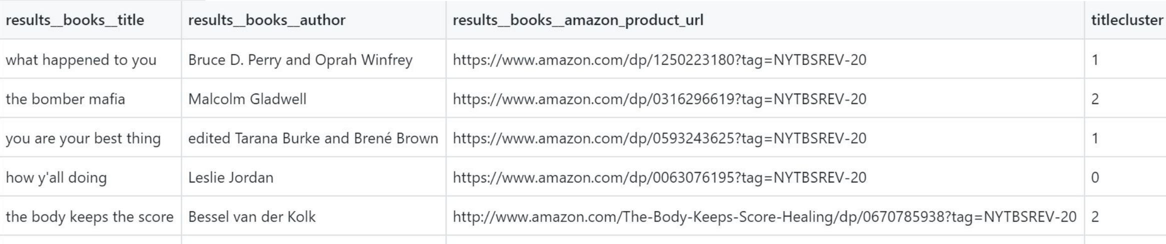
Title Analysis: Clustering (# with elbow method)

NYT + GBooks



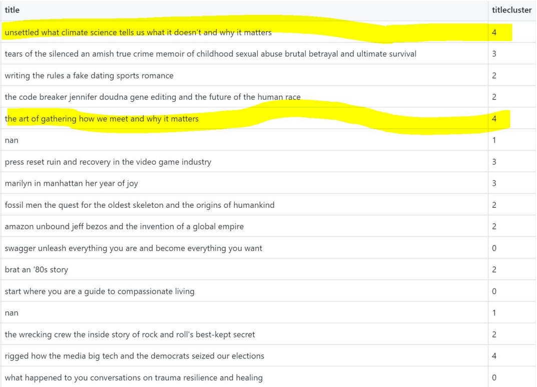
title	titlecluster
think again the power of knowing what you don't know	0
nan	3
crying in h mart a memoir	4
how y'all doing misadventures and mischief from a life well lived	2
empire of pain the secret history of the sackler dynasty	1
the code breaker jennifer doudna gene editing and the future of the human race	1
what happened to you conversations on trauma resilience and healing	0
antitrust taking on monopoly power from the gilded age to the digital age	1

NYT ONLY (6)



results__books__title	results__books__author	results__books__amazon_product_url	titlecluster
what happened to you	Bruce D. Perry and Oprah Winfrey	https://www.amazon.com/dp/1250223180?tag=NYTBSREV-20	1
the bomber mafia	Malcolm Gladwell	https://www.amazon.com/dp/0316296619?tag=NYTBSREV-20	2
you are your best thing	edited Tarana Burke and Brené Brown	https://www.amazon.com/dp/0593243625?tag=NYTBSREV-20	1
how y'all doing	Leslie Jordan	https://www.amazon.com/dp/0063076195?tag=NYTBSREV-20	0
the body keeps the score	Bessel van der Kolk	http://www.amazon.com/The-Body-Keeps-Score-Healing/dp/0670785938?tag=NYTBSREV-20	2

AMAZON(5)

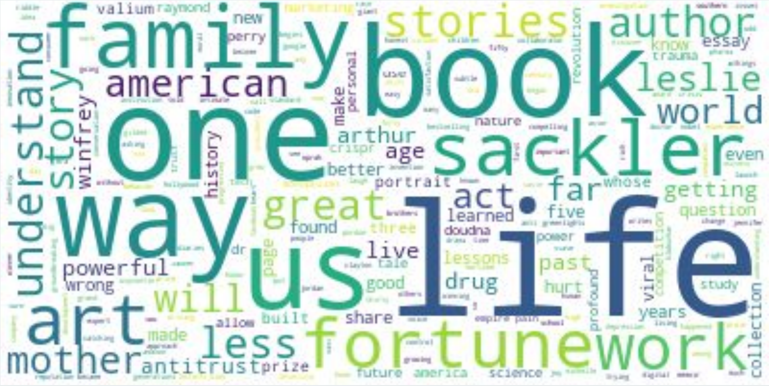


title	titlecluster
unsettled what climate science tells us what it doesn't and why it matters	4
tears of the silenced an amish true crime memoir of childhood sexual abuse brutal betrayal and ultimate survival	3
writing the rules a fake dating sports romance	2
the code breaker jennifer doudna gene editing and the future of the human race	2
the art of gathering how we meet and why it matters	4
nan	1
press reset ruin and recovery in the video game industry	3
marilyn in manhattan her year of joy	3
fossil men the quest for the oldest skeleton and the origins of humankind	2
amazon unbound jeff bezos and the invention of a global empire	2
swagger unleash everything you are and become everything you want	0
brat an '80s story	2
start where you are a guide to compassionate living	0
nan	1
the wrecking crew the inside story of rock and roll's best-kept secret	2
rigged how the media big tech and the democrats seized our elections	4
what happened to you conversations on trauma resilience and healing	0

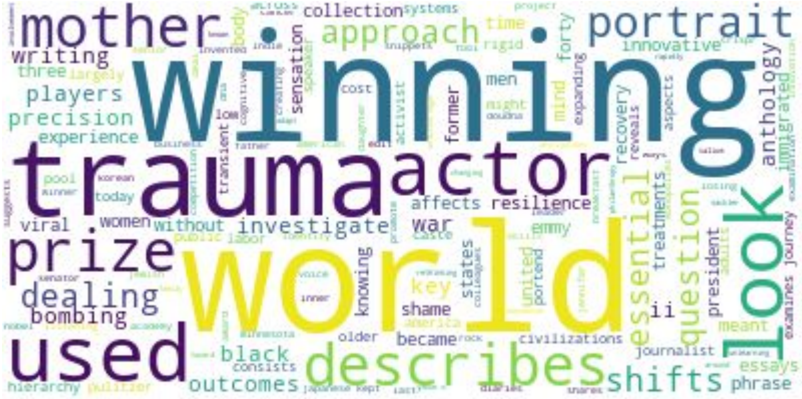
Description Text Analysis

Description Analysis: Exploratory Data Analysis using Wordcloud

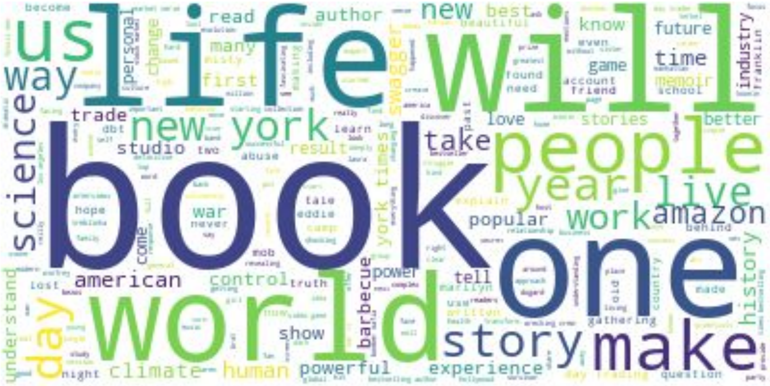
NYT + GBOOKS



NYT ONLY

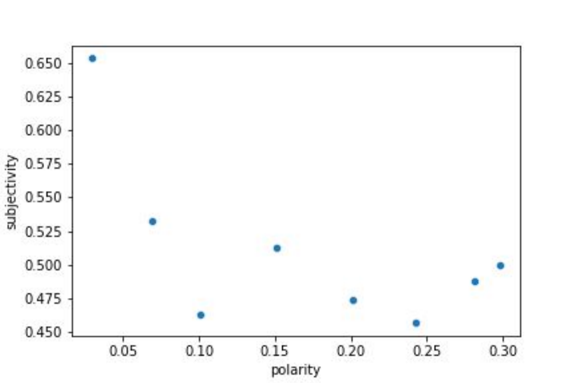


AMAZON

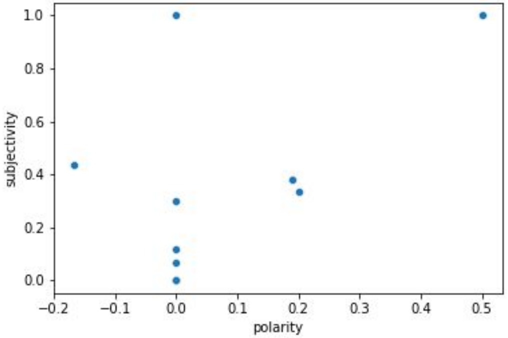


Description Analysis: Textblob

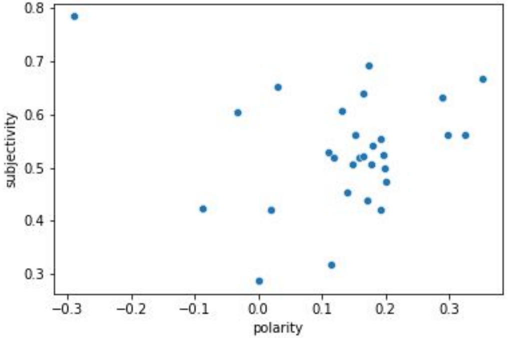
NYT + GBooks



NYT ONLY

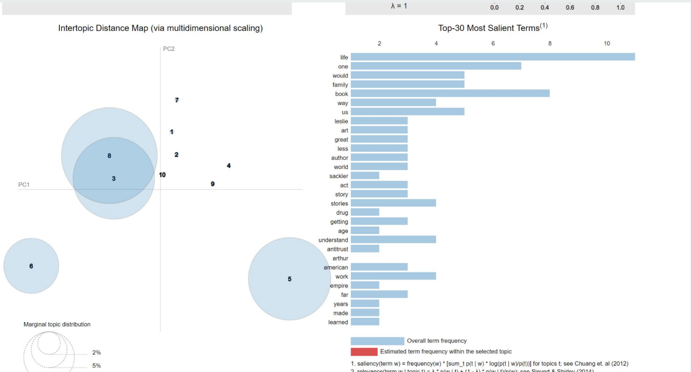


AMAZON



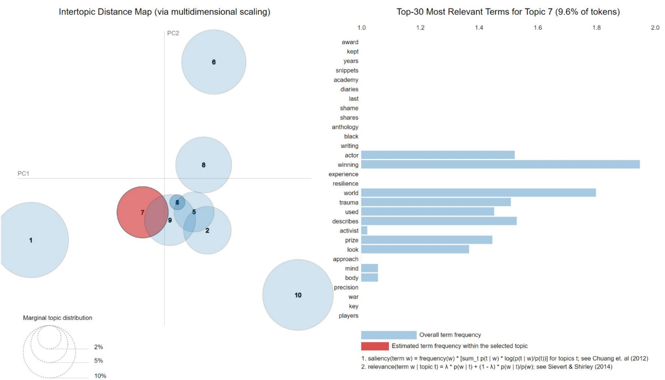
Description Analysis: LDA Topic Modeling (visualized via pyLDAvis)

NYT + Gbooks



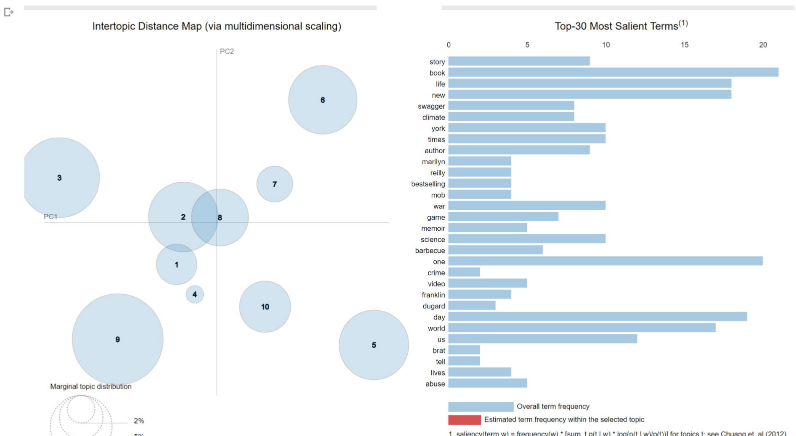
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

NYT ONLY



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

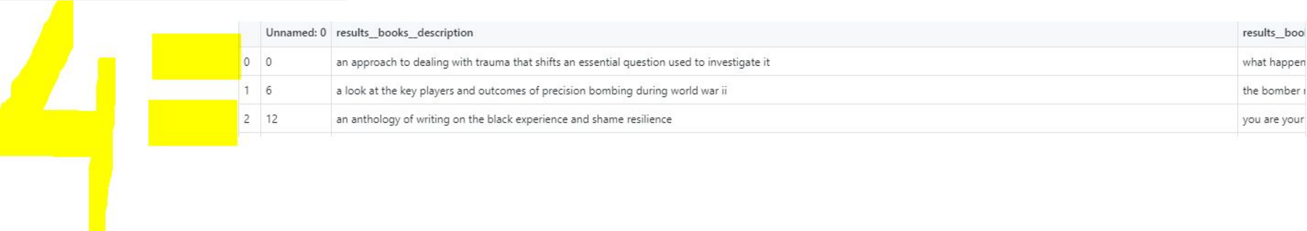
AMAZON ONLY



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

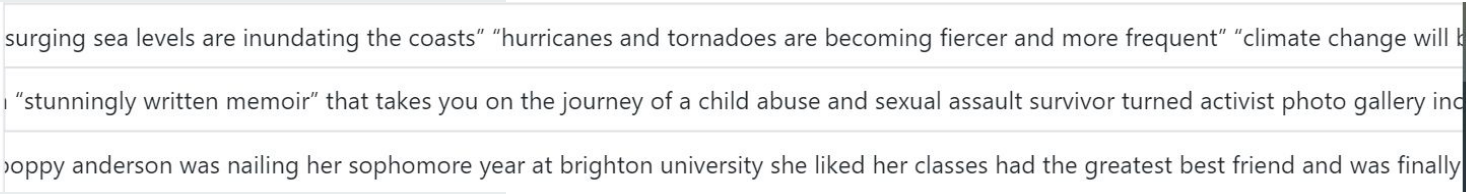
Description Analysis: Clustering (# with elbow method)

NYT ONLY (6)



	Unnamed: 0	results_books_description	results_boo
0	0	an approach to dealing with trauma that shifts an essential question used to investigate it	what happen
1	6	a look at the key players and outcomes of precision bombing during world war ii	the bomber
2	12	an anthology of writing on the black experience and shame resilience	you are your

AMAZON ONLY (8)

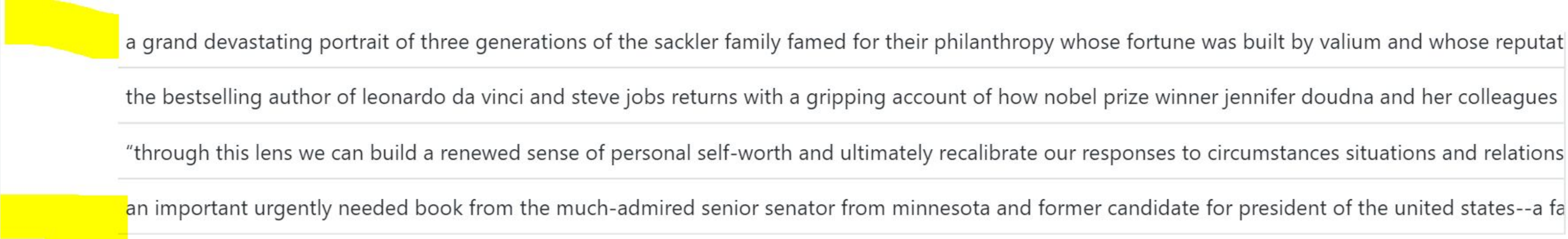


surging sea levels are inundating the coasts" "hurricanes and tornadoes are becoming fiercer and more frequent" "climate change will b

"stunningly written memoir" that takes you on the journey of a child abuse and sexual assault survivor turned activist photo gallery inc

poppy anderson was nailing her sophomore year at brighton university she liked her classes had the greatest best friend and was finally

NYT GBOOKS (6)



a grand devastating portrait of three generations of the sackler family famed for their philanthropy whose fortune was built by valium and whose reputat

the bestselling author of leonardo da vinci and steve jobs returns with a gripping account of how nobel prize winner jennifer doudna and her colleagues

"through this lens we can build a renewed sense of personal self-worth and ultimately recalibrate our responses to circumstances situations and relations

an important urgently needed book from the much-admired senior senator from minnesota and former candidate for president of the united states--a fa

Conclusion

Thoughts on Data Analysis (Suggestions welcome!)

- **WORDCLOUD**

  Could potentially be replaced by better data visualizations/only as a preliminary step in EDA.

- **TEXTBLOB**

  Might be better used for reviews (although it would be interesting to see sentiment analysis over a large dataset).

- **LDA TOPIC MODELING**

  Will continue to use as done here and will only get even better with larger planned dataset. Will work on making code less "out of box" and more useful

- **CLUSTERING**

  Same as above but need to figure out better visualization method. Will work on making code less "out of box" and more useful

Plans for Next Week

- Work with 1 year dataset to come up with more useful findings.
- Implement suggestions.
- Optimize how I work with Google Books API - how to make sure only one result for as many books as possible.