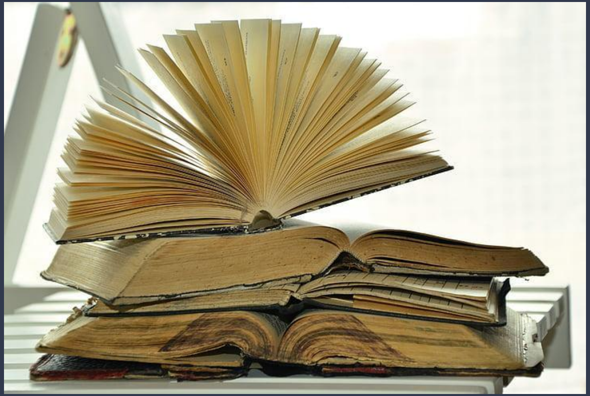
Cohesive Topics for Books

Using TF-IDF, Word2Vec, and NMF to cluster books by description



Dataset

- Amazon book dataset
- 212404 objects, 10 features

Features:

- title
- description
- authors
- image
- previewLink
- publisher
- publishedDate
- infoLink
- categories
- ratingsCount

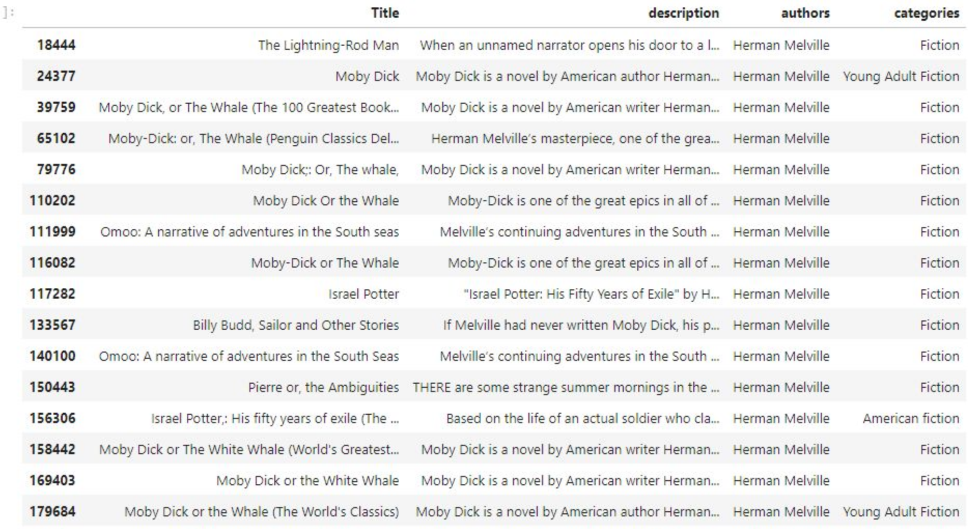
Purpose

- Problem:
    - Dataset contains 1868 unique topics
    - Only 122 categories contain 5+ books
    - Large discrepancy between membership in most and least frequent categories

- Solution:
    - Use unsupervised learning to group books into fewer topics
    - Use analysis of book descriptions to group similar books
- Tools:
    - TF-IDF
    - Word2Vec
    - NMF

Cleaning

- Delete columns:
  - Publisher information
  - Links
  - Ratings count
- Remove multiple editions of same book
- Remove books with descriptions below median length
  - Median length: 553 characters

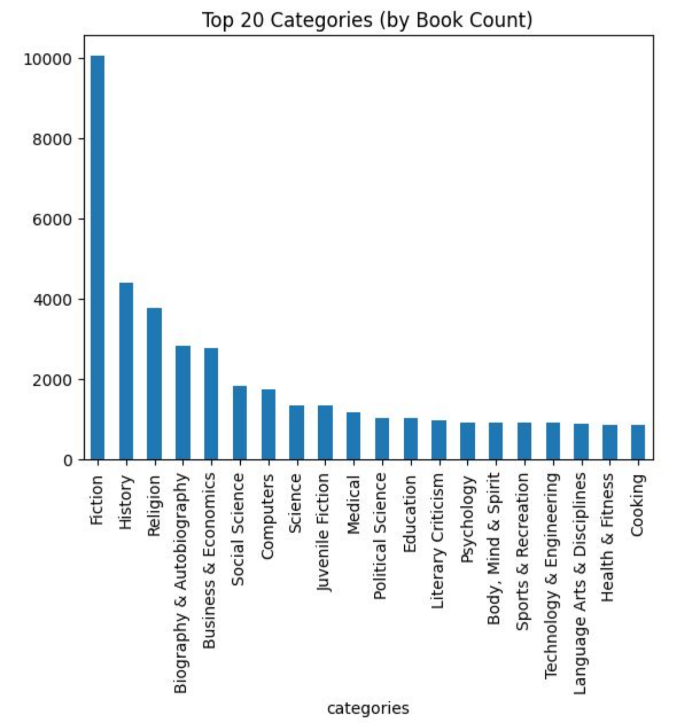


]:

	Title	description	authors	categories
18444	The Lightning-Rod Man	When an unnamed narrator opens his door to a l...	Herman Melville	Fiction
24377	Moby Dick	Moby Dick is a novel by American author Herman...	Herman Melville	Young Adult Fiction
39759	Moby Dick, or The Whale (The 100 Greatest Book...	Moby Dick is a novel by American writer Herman...	Herman Melville	Fiction
65102	Moby-Dick: or, The Whale (Penguin Classics Del...	Herman Melville's masterpiece, one of the grea...	Herman Melville	Fiction
79776	Moby Dick;: Or, The whale,	Moby Dick is a novel by American writer Herman...	Herman Melville	Fiction
110202	Moby Dick Or the Whale	Moby-Dick is one of the great epics in all of ...	Herman Melville	Fiction
111999	Omoo: A narrative of adventures in the South seas	Melville's continuing adventures in the South ...	Herman Melville	Fiction
116082	Moby-Dick or The Whale	Moby-Dick is one of the great epics in all of ...	Herman Melville	Fiction
117282	Israel Potter	"Israel Potter: His Fifty Years of Exile" by H...	Herman Melville	Fiction
133567	Billy Budd, Sailor and Other Stories	If Melville had never written Moby Dick, his p...	Herman Melville	Fiction
140100	Omoo: A narrative of adventures in the South Seas	Melville's continuing adventures in the South ...	Herman Melville	Fiction
150443	Pierre or, the Ambiguities	THERE are some strange summer mornings in the ...	Herman Melville	Fiction
156306	Israel Potter,: His fifty years of exile (The ...	Based on the life of an actual soldier who cla...	Herman Melville	American fiction
158442	Moby Dick or The White Whale (World's Greatest...	Moby Dick is a novel by American writer Herman...	Herman Melville	Fiction
169403	Moby Dick or the White Whale	Moby Dick is a novel by American writer Herman...	Herman Melville	Fiction
179684	Moby Dick or the Whale (The World's Classics)	Moby Dick is a novel by American author Herman...	Herman Melville	Young Adult Fiction

- 9 editions of Moby Dick were present in original dataset
- Multiple editions would overrepresent certain books in the final topics

EDA

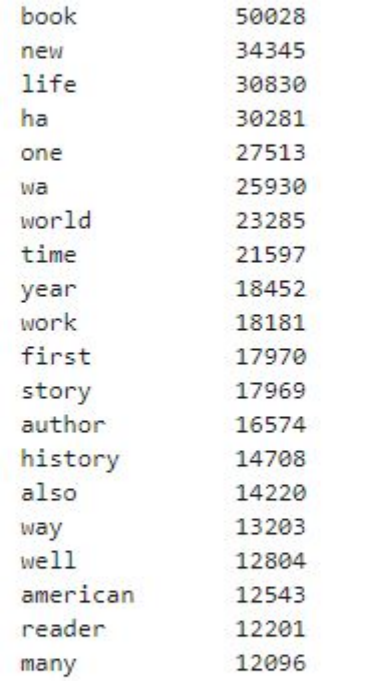


- Fiction is the most frequent category (17.9% of books are categorized as fiction)
- The top 5 categories account for 42.4% of books
- 1396 categories have only one book

Metric	Books in Category
Mean	30.1
Median	1.0
Min	1
Max	10048

EDA

20 Most Frequent Words



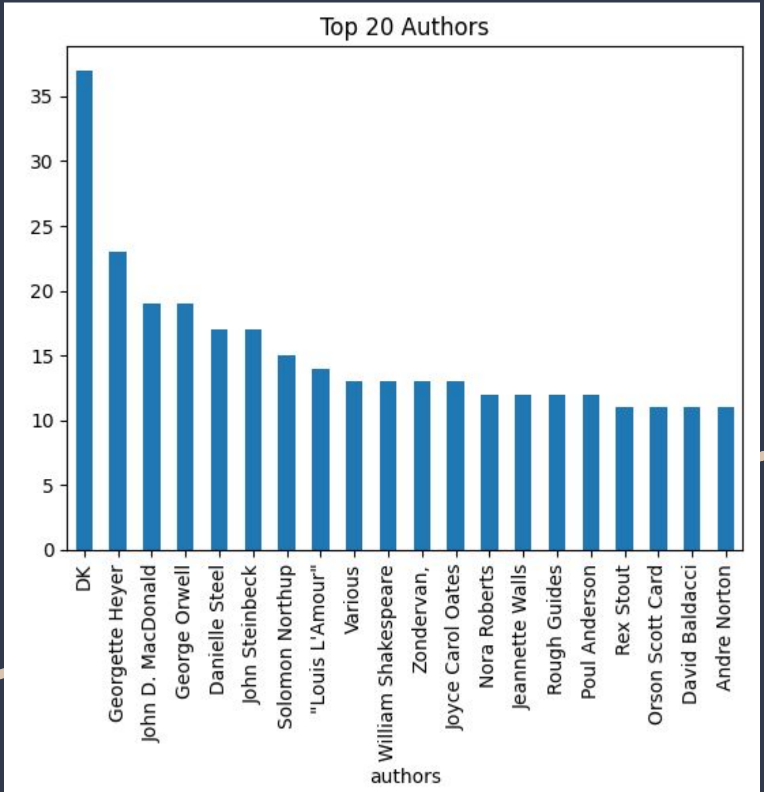
```
book        50028
new         34345
life        30830
ha          30281
one         27513
wa          25930
world       23285
time        21597
year        18452
work        18181
first       17970
story       17969
author      16574
history     14708
also        14220
way         13203
well        12804
american    12543
reader      12201
many        12096
```

- 147,802 unique words in the book descriptions
- 1,183 words occur over 1000 times
- 67,679 words occur only once

Description Metrics

Metric	Value (in characters)
Mean	939.2
Median	834.0
Min	553
Max	14086

EDA

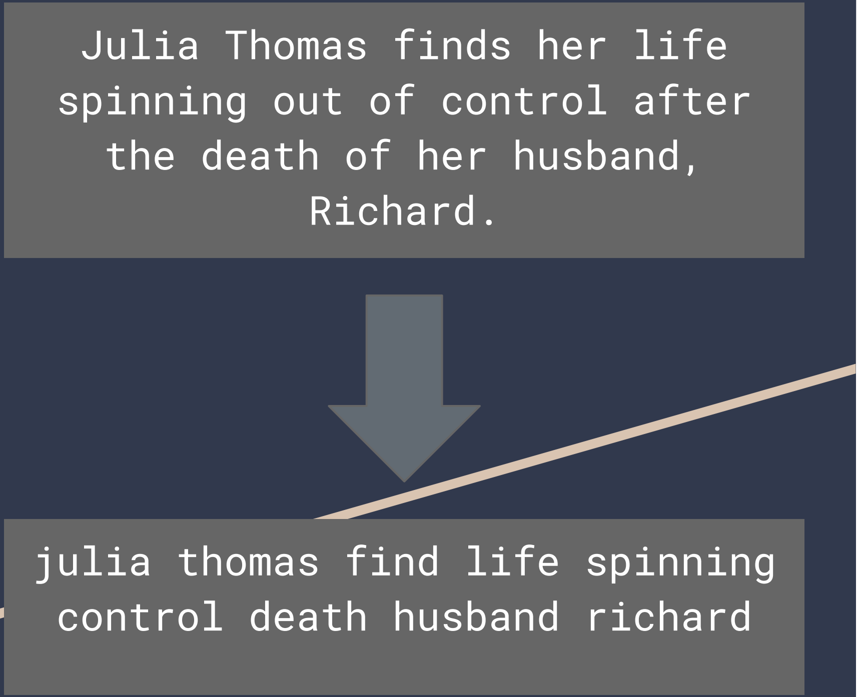


- Dataset contains 35,384 unique authors
- DK is the most frequent with 37 books

Author Metrics

Metric	Books per Author
Mean	1.2
Median	1.0
Min	1
Max	37

Data preprocessing



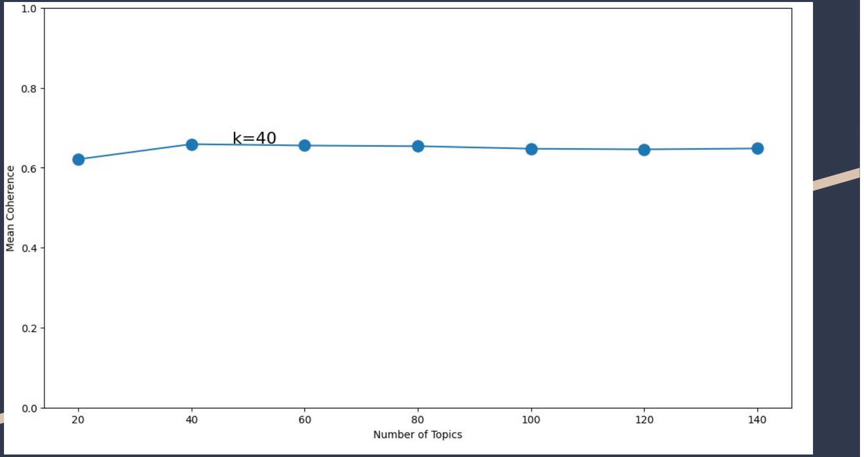
Julia Thomas finds her life spinning out of control after the death of her husband, Richard.

julia thomas find life spinning control death husband richard

- Cleaned description text:
  - Remove punctuation, stop words, and non-alphabetic characters
  - Lemmatized words
  - Converted to lowercase
- Used TF-IDF to calculate importance of words within the descriptions
- Used Word2Vec determine associations between words

Model building: NMF

Mean Coherence for k Components



- Used random subsets for reasonable processing time
- Evaluation metric: coherence within topics
- Coherence is relatively stable across components between 40 and 140
- 40 is selected as optimal because it is the simplest model with highest coherence

k	Mean Coherence
20	0.62
40	0.66
60	0.66
80	0.65
100	0.65
120	0.65
140	0.65

Tuned other hyperparameters

- Compared:
    - Init: ‘nndsvd,’ ‘nndsvdar’
    - Solver: ‘cd,’ ‘mu’
    - Beta_loss: ‘frobenius,’ ‘kullback-leibler’

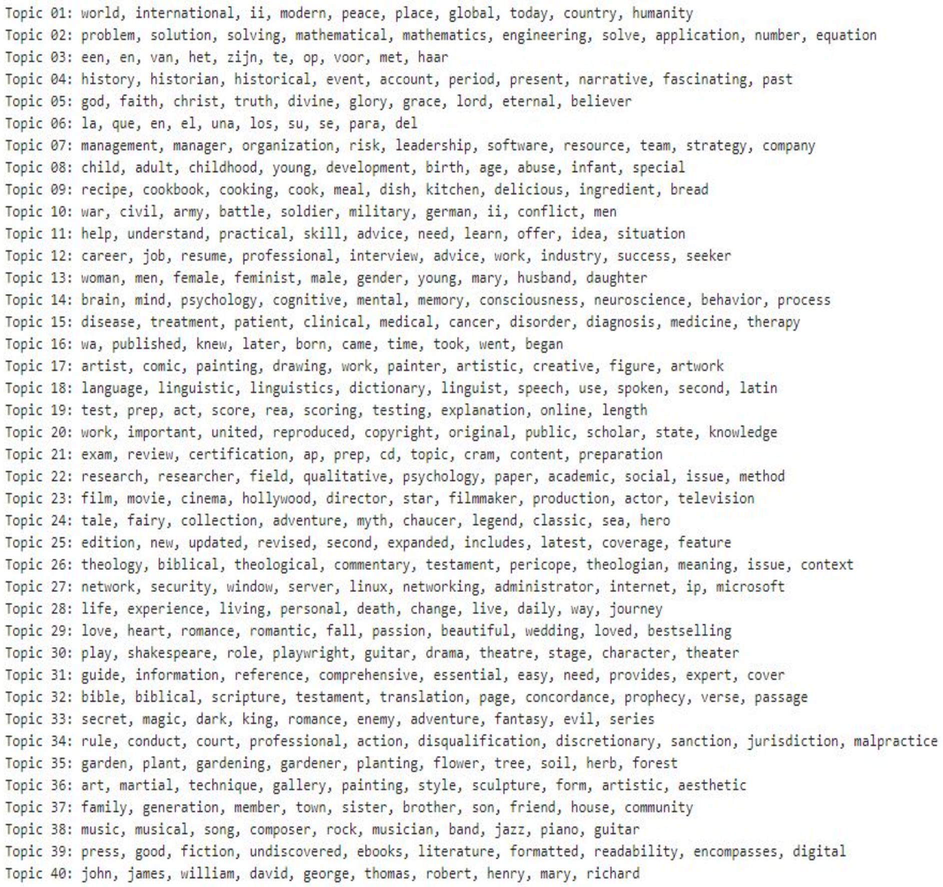
Optimal model:

- Solver = ‘cd’
- Init = ‘nndsvd’
- Beta_loss = ‘frobenius’

Mean coherence across 5 random subsets: 0.65

Produced Topics

- Keywords for each topic can be used to label each topic
- Examples:
  - Topic 9: Cooking
  - Topic 29: Romance
  - Topic 38: Music



```
Topic 01: world, international, ii, modern, peace, place, global, today, country, humanity
Topic 02: problem, solution, solving, mathematical, mathematics, engineering, solve, application, number, equation
Topic 03: een, en, van, het, zijn, te, op, voor, met, haar
Topic 04: history, historian, historical, event, account, period, present, narrative, fascinating, past
Topic 05: god, faith, christ, truth, divine, glory, grace, lord, eternal, believer
Topic 06: la, que, en, el, una, los, su, se, para, del
Topic 07: management, manager, organization, risk, leadership, software, resource, team, strategy, company
Topic 08: child, adult, childhood, young, development, birth, age, abuse, infant, special
Topic 09: recipe, cookbook, cooking, cook, meal, dish, kitchen, delicious, ingredient, bread
Topic 10: war, civil, army, battle, soldier, military, german, ii, conflict, men
Topic 11: help, understand, practical, skill, advice, need, learn, offer, idea, situation
Topic 12: career, job, resume, professional, interview, advice, work, industry, success, seeker
Topic 13: woman, men, female, feminist, male, gender, young, mary, husband, daughter
Topic 14: brain, mind, psychology, cognitive, mental, memory, consciousness, neuroscience, behavior, process
Topic 15: disease, treatment, patient, clinical, medical, cancer, disorder, diagnosis, medicine, therapy
Topic 16: wa, published, knew, later, born, came, time, took, went, began
Topic 17: artist, comic, painting, drawing, work, painter, artistic, creative, figure, artwork
Topic 18: language, linguistic, linguistics, dictionary, linguist, speech, use, spoken, second, latin
Topic 19: test, prep, act, score, rea, scoring, testing, explanation, online, length
Topic 20: work, important, united, reproduced, copyright, original, public, scholar, state, knowledge
Topic 21: exam, review, certification, ap, prep, cd, topic, cram, content, preparation
Topic 22: research, researcher, field, qualitative, psychology, paper, academic, social, issue, method
Topic 23: film, movie, cinema, hollywood, director, star, filmmaker, production, actor, television
Topic 24: tale, fairy, collection, adventure, myth, chaucer, legend, classic, sea, hero
Topic 25: edition, new, updated, revised, second, expanded, includes, latest, coverage, feature
Topic 26: theology, biblical, theological, commentary, testament, pericope, theologian, meaning, issue, context
Topic 27: network, security, window, server, linux, networking, administrator, internet, ip, microsoft
Topic 28: life, experience, living, personal, death, change, live, daily, way, journey
Topic 29: love, heart, romance, romantic, fall, passion, beautiful, wedding, loved, bestselling
Topic 30: play, shakespeare, role, playwright, guitar, drama, theatre, stage, character, theater
Topic 31: guide, information, reference, comprehensive, essential, easy, need, provides, expert, cover
Topic 32: bible, biblical, scripture, testament, translation, page, concordance, prophecy, verse, passage
Topic 33: secret, magic, dark, king, romance, enemy, adventure, fantasy, evil, series
Topic 34: rule, conduct, court, professional, action, disqualification, discretionary, sanction, jurisdiction, malpractice
Topic 35: garden, plant, gardening, gardener, planting, flower, tree, soil, herb, forest
Topic 36: art, martial, technique, gallery, painting, style, sculpture, form, artistic, aesthetic
Topic 37: family, generation, member, town, sister, brother, son, friend, house, community
Topic 38: music, musical, song, composer, rock, musician, band, jazz, piano, guitar
Topic 39: press, good, fiction, undiscovered, ebooks, literature, formatted, readability, encompasses, digital
Topic 40: john, james, william, david, george, thomas, robert, henry, mary, richard
```

Discussion and Future directions

- Model produced 40 cohesive topics
- Keywords indicate meaningful clustering
- Significant reduction from 1868 categories

- Limit dataset to English books
  - Model did group foreign language books together (Topics 3 and 6), but was intended for English-language books
  - Performance may increase
- Incorporate other features into analysis:
  - Image analysis of book covers

References

Akdogan, A. (2021, Jul 22). *Word Embedding Techniques: Word2Vec and TF-IDF Explained.* Medium. https://towardsdatascience.com/word-embedding-techniques-word2vec-and-tf-idf-explained-c5d02e34d08

Andreichuk, V. (2023, Dec 10). *Non-negative Matrix Factorization (NMF) for the Grouping of Articles' Titles* . Medium. https://medium.com/@vlad.andreichuk/non-negative-matrix-factorization-nmf-for-the-grouping-of-articles-titles-a73e654b6244

Bekheet, M. (2022, September 13). *Amazon Books Reviews*. Kaggle. https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews/

R, Y. (2024, Jan 2). *Python Lemmatization with NLTK*. Geeks for Geeks. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/