CSU | Cleveland State University

# Book Recommendation System

Handong Wang 2874116

Vijay Rajkumar Yadav 2872733

Swapnanil Das 2885570

- Motivation & Background
- Data Cleaning & Exploration
- Calculate Book Similarity
  - **Word2Vec** Based Book Similarity
  - **TF-IDF** Based Book Similarity
- Generating recommendations
- Conclusions & Future Work

goodreads

goodreads CHOICE AWARDS 2024

Announcing *Readers' Favorite* Books of 2024

Deciding what to read next?

You're in the right place. Tell us what titles or genres you've enjoyed in the past, and we'll give you surprisingly insightful recommendations.

What are your friends reading?

Chances are your friends are discussing their favorite (and least favorite) books on Goodreads.

What will *you* discover?

Because Deborah liked...

She discovered:

Steam Punk, Science Fiction, Fantasy

Because Brian liked...

He discovered:

Decision-making, Sociology, Marketing

BONESHAKER
TIM POWERS THE ANUBIS GATES
CLEMENTINE
MIKE RESNICK THE BUNTLINE SPECIAL
INFERNAL DEVICES K.W. JETER
Nudge
TRAFFIC
PREDICTABLY IRRATIONAL
THINKING, FAST AND SLOW DANIEL KAHNEMAN

CSU Cleveland State University

# Background

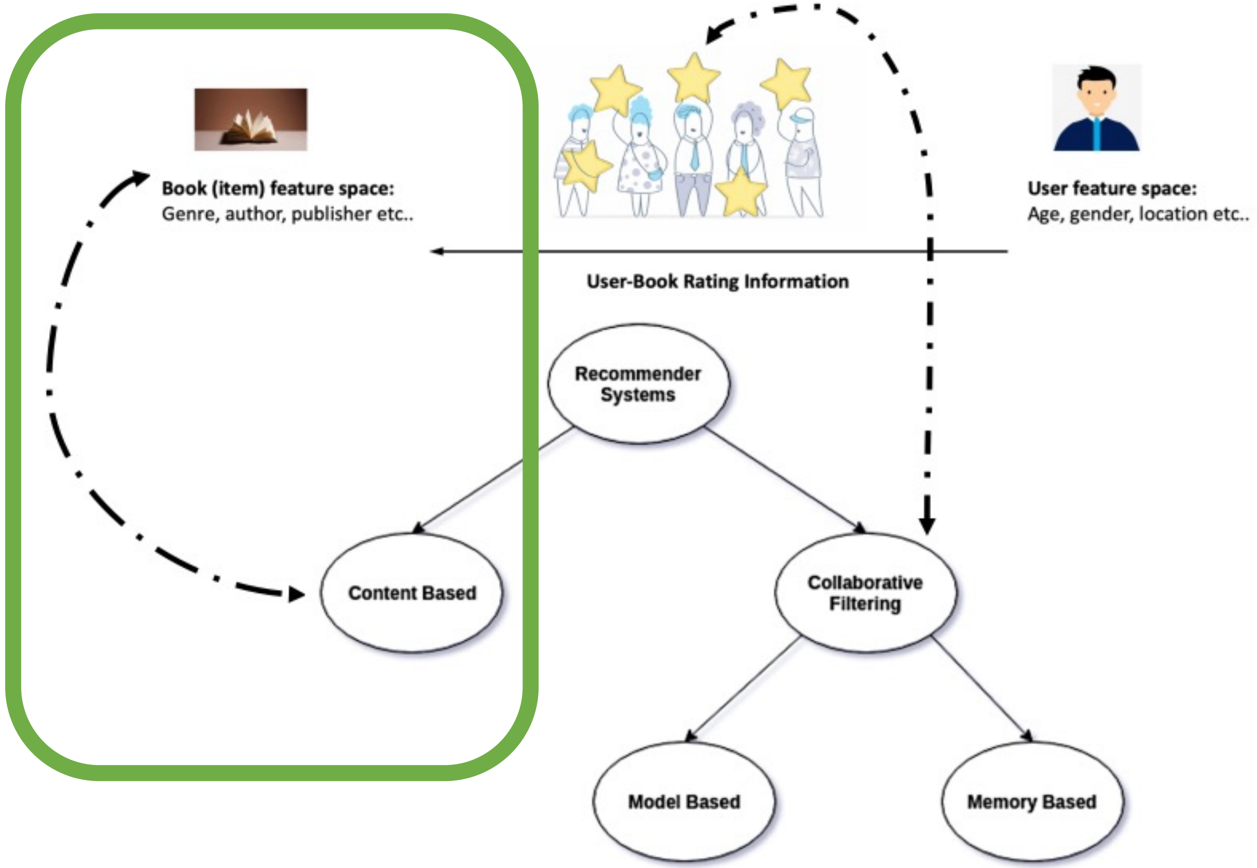
- For Users: Users who like books often like books by the same author or on the same subject. In order to make it easier for users to find books they like.

- For Company: In order to improve the competitiveness of the company's products, it is necessary to increase the user's immersion time. If users find more books they like through book recommendations, it can increase user time and improve product traffic data.

# Motivation

***In this project, we focus on content-based recommendations using NLP techniques.***

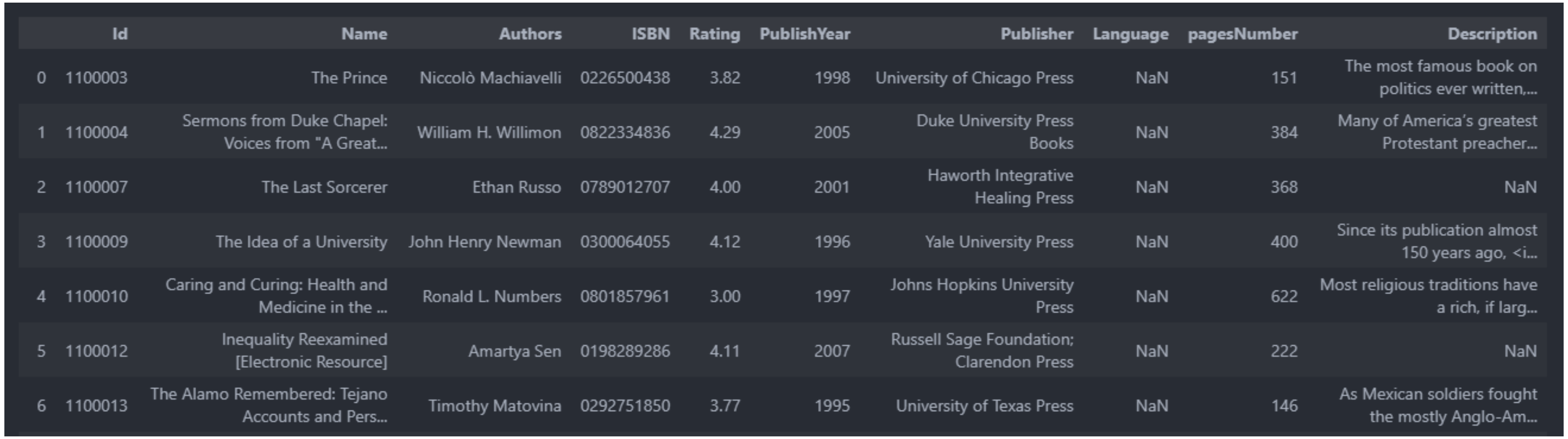
**Content based recommendations** based on user past likes & dislikes – System recommends items similar to items users have liked based on item feature space


Book (item) feature space:
Genre, author, publisher etc..
User feature space:
Age, gender, location etc..
User-Book Rating Information
Recommender Systems
Content Based
Collaborative Filtering
Model Based
Memory Based

CSU | Cleveland State University

# Data Cleaning & Exploration

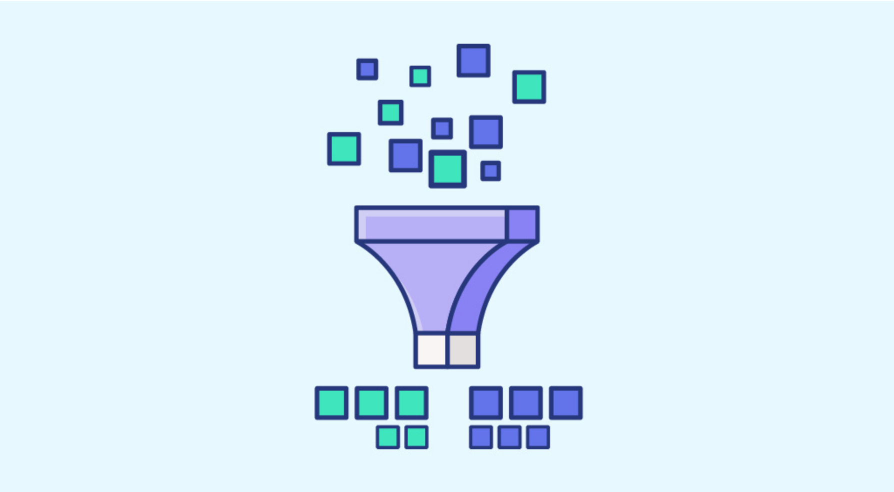
- **Data Source**
  - It is a Goodreads book dataset containing book details and user rating of 10M books.
  - In This Project, We only use **5000** books.
  - It contains information such as book name, authors, publishers, publishing year, rating, description, review count, page number etc.



|  | Id | Name | Authors | ISBN | Rating | PublishYear | Publisher | Language | pagesNumber | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1100003 | The Prince | Niccolò Machiavelli | 0226500438 | 3.82 | 1998 | University of Chicago Press | NaN | 151 | The most famous book on politics ever written,... |
| 1 | 1100004 | Sermons from Duke Chapel: Voices from "A Great... | William H. Willimon | 0822334836 | 4.29 | 2005 | Duke University Press Books | NaN | 384 | Many of America's greatest Protestant preacher... |
| 2 | 1100007 | The Last Sorcerer | Ethan Russo | 0789012707 | 4.00 | 2001 | Haworth Integrative Healing Press | NaN | 368 | NaN |
| 3 | 1100009 | The Idea of a University | John Henry Newman | 0300064055 | 4.12 | 1996 | Yale University Press | NaN | 400 | Since its publication almost 150 years ago, <i... |
| 4 | 1100010 | Caring and Curing: Health and Medicine in the ... | Ronald L. Numbers | 0801857961 | 3.00 | 1997 | Johns Hopkins University Press | NaN | 622 | Most religious traditions have a rich, if larg... |
| 5 | 1100012 | Inequality Reexamined [Electronic Resource] | Amartya Sen | 0198289286 | 4.11 | 2007 | Russell Sage Foundation; Clarendon Press | NaN | 222 | NaN |
| 6 | 1100013 | The Alamo Remembered: Tejano Accounts and Pers... | Timothy Matovina | 0292751850 | 3.77 | 1995 | University of Texas Press | NaN | 146 | As Mexican soldiers fought the mostly Anglo-Am... |

# Data Cleaning & Exploration

- Data Cleaning
  - Data cleaning in NLP involves transforming textual data into a machine-readable format to reduce model complexity and enhance accuracy. This process helps avoid processing irrelevant words and ensures that the model treats the same words equally, regardless of punctuation or letter case.
  - **Feature Selection**
  - **Data Refinement**
    - **Remove URLs and HTML tags from description**
    - **Eliminate NaN values**
    - **Exclude missing data**



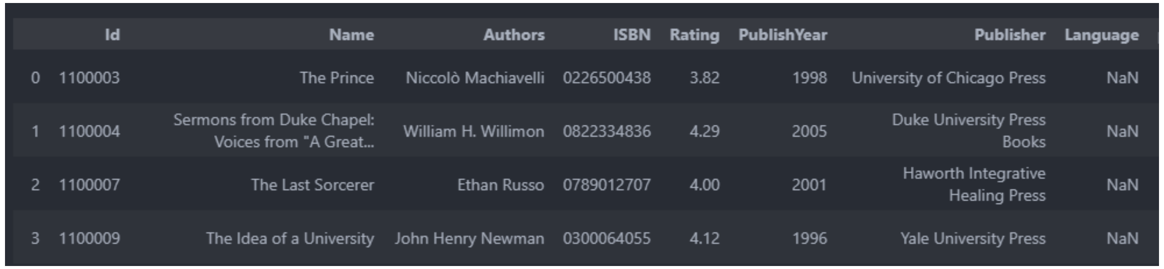
# Data Cleaning & Exploration

- Feature Selection

  - **ISBN is Not Usable:** The ISBN doesn't provide useful content information and adds unnecessary complexity to the model.

  - **Language Has Too Many NaN Values**: The language feature has too many missing values, which could lead to unreliable predictions. It's better to exclude it.

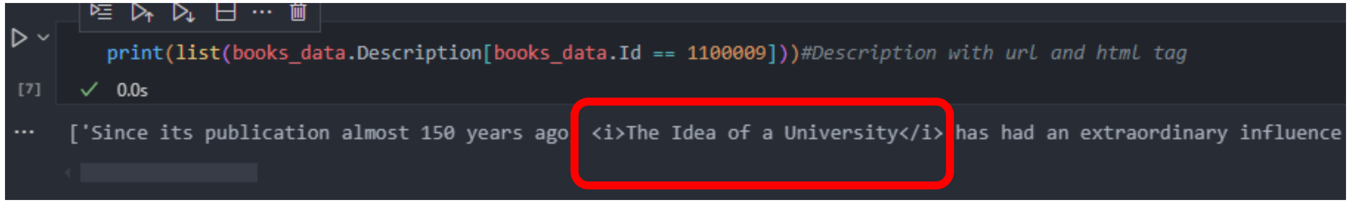
  - Final Features:
    - `Id, Name, Authors, PublishYear, Publisher, Description, Rating, pagesNumber`



| | Id | Name | Authors | ISBN | Rating | PublishYear | Publisher | Language |
|---|---|---|---|---|---|---|---|---|
| 0 | 1100003 | The Prince | Niccolò Machiavelli | 0226500438 | 3.82 | 1998 | University of Chicago Press | NaN |
| 1 | 1100004 | Sermons from Duke Chapel: Voices from "A Great... | William H. Willimon | 0822334836 | 4.29 | 2005 | Duke University Press Books | NaN |
| 2 | 1100007 | The Last Sorcerer | Ethan Russo | 0789012707 | 4.00 | 2001 | Haworth Integrative Healing Press | NaN |
| 3 | 1100009 | The Idea of a University | John Henry Newman | 0300064055 | 4.12 | 1996 | Yale University Press | NaN |

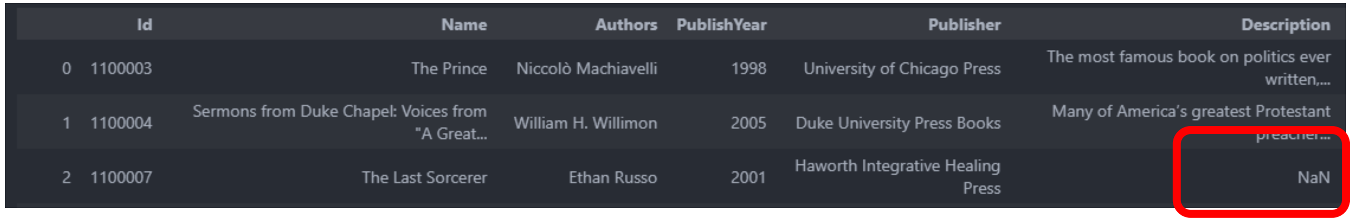
# Data Cleaning & Exploration

- **Data Refinement**
  - **Remove URLs and HTML tags from description**



```
print(list(books_data.Description[books_data.Id == 1100009]))#Description with url and html tag
[7]  ✓ 0.0s
...  ['Since its publication almost 150 years ago <i>The Idea of a University</i> has had an extraordinary influence
```

  - **Eliminate NaN values**

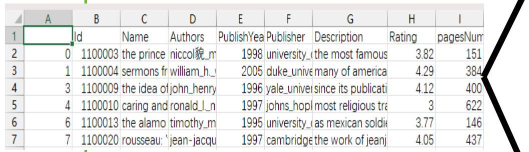


| | Id | Name | Authors | PublishYear | Publisher | Description |
|---|---|---|---|---|---|---|
| 0 | 1100003 | The Prince | Niccolò Machiavelli | 1998 | University of Chicago Press | The most famous book on politics ever written,... |
| 1 | 1100004 | Sermons from Duke Chapel: Voices from "A Great... | William H. Willimon | 2005 | Duke University Press Books | Many of America's greatest Protestant preacher... |
| 2 | 1100007 | The Last Sorcerer | Ethan Russo | 2001 | Haworth Integrative Healing Press | NaN |

CSU | Cleveland State University

# Calculate Book Similarity

- **Pipeline**

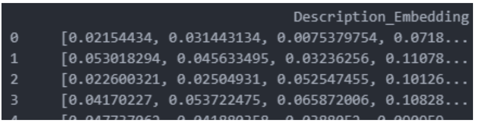


|   | Id | Name | Authors | PublishYea | Publisher | Description | Rating | pagesNum |
|---|---|---|---|---|---|---|---|---|
| 0 | 1100003 | the prince | niccol貌_m | 1998 | university_( | the most famous | 3.82 | 151 |
| 1 | 1100004 | sermons fr | william_h._ | 2005 | duke_unive | many of america | 4.29 | 384 |
| 3 | 1100009 | the idea of | john_henry | 1996 | yale_univer | since its publicati | 4.12 | 400 |
| 4 | 1100010 | caring and | ronald_l._n | 1997 | johns_hopl | most religious tra | 3 | 622 |
| 6 | 1100013 | the alamo | timothy_m | 1995 | university_( | as mexican soldie | 3.77 | 146 |
| 7 | 1100020 | rousseau: ' | jean-jacqu | 1997 | cambridge | the work of jeanj | 4.05 | 437 |

Processed Book Data

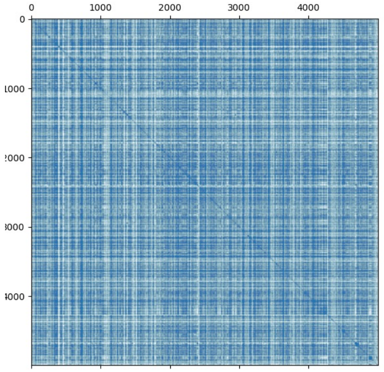
Description

Word2Vec Model



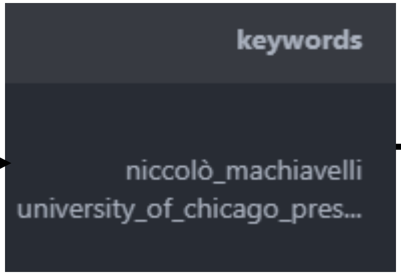
```
                                      Description_Embedding
0    [0.02154434, 0.031443134, 0.0075379754, 0.0718...
1    [0.053018294, 0.045633495, 0.03236256, 0.11078...
2    [0.022600321, 0.02504931, 0.052547455, 0.10126...
3    [0.04170227, 0.053722475, 0.065872006, 0.10828...
```

Description Embedding



Similarity Matrix By **Word2Vec**

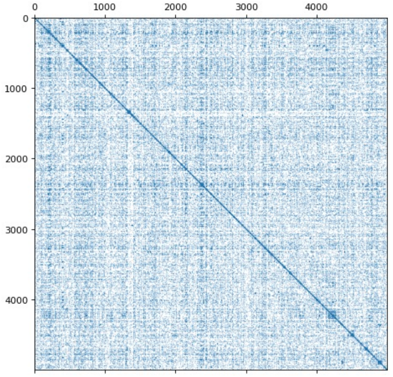
KeyBERT Model



```
                keywords

     niccolò_machiavelli
university_of_chicago_pres...
```

Key Words

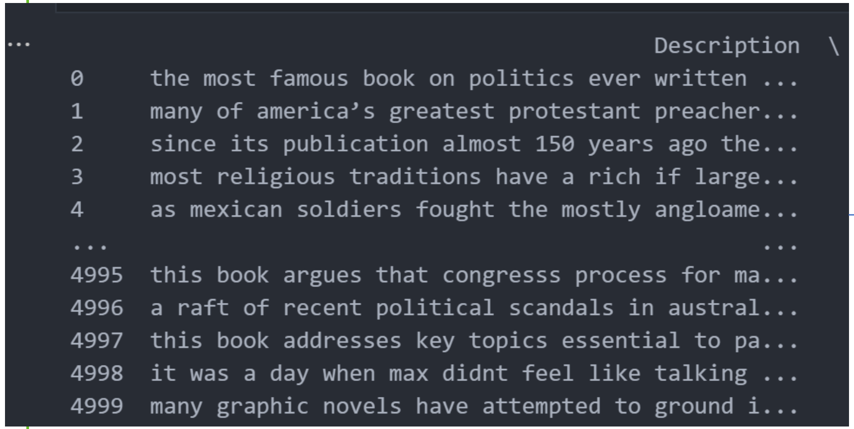
TF-IDF



Similarity Matrix By **TF-IDF**

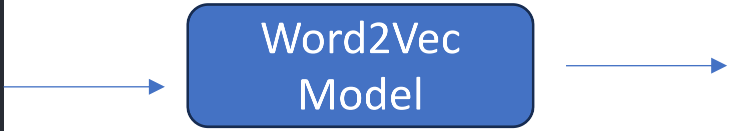
CSU | Cleveland State University

# Word2Vec Based Book Similarity

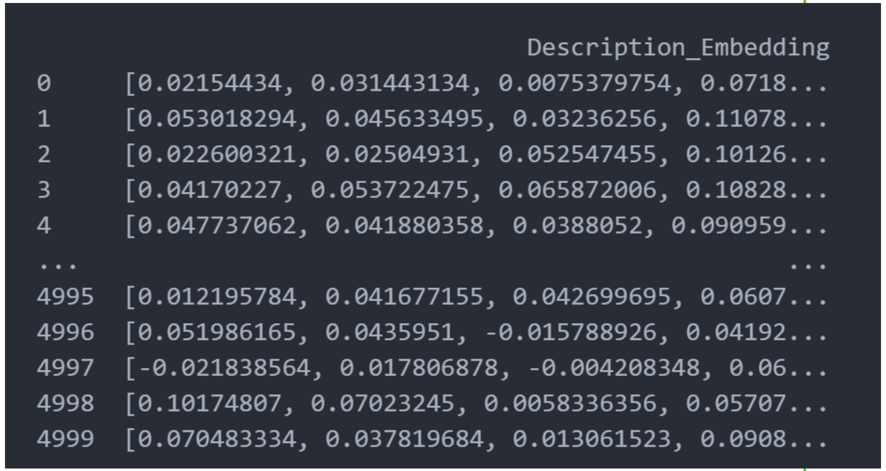


```
...                                               Description  \
0     the most famous book on politics ever written ...
1     many of america’s greatest protestant preacher...
2     since its publication almost 150 years ago the...
3     most religious traditions have a rich if large...
4     as mexican soldiers fought the mostly angloame...
...                                                 ...
4995  this book argues that congresss process for ma...
4996  a raft of recent political scandals in austral...
4997  this book addresses key topics essential to pa...
4998  it was a day when max didnt feel like talking ...
4999  many graphic novels have attempted to ground i...
```

Processed Book Data Description



Word2Vec Model



```
                                  Description_Embedding
0     [0.02154434, 0.031443134, 0.0075379754, 0.0718...
1     [0.053018294, 0.045633495, 0.03236256, 0.11078...
2     [0.022600321, 0.02504931, 0.052547455, 0.10126...
3     [0.04170227, 0.053722475, 0.065872006, 0.10828...
4     [0.047737062, 0.041880358, 0.0388052, 0.090959...
...                                                 ...
4995  [0.012195784, 0.041677155, 0.042699695, 0.0607...
4996  [0.051986165, 0.0435951, -0.015788926, 0.04192...
4997  [-0.021838564, 0.017806878, -0.004208348, 0.06...
4998  [0.10174807, 0.07023245, 0.0058336356, 0.05707...
4999  [0.070483334, 0.037819684, 0.013061523, 0.0908...
```

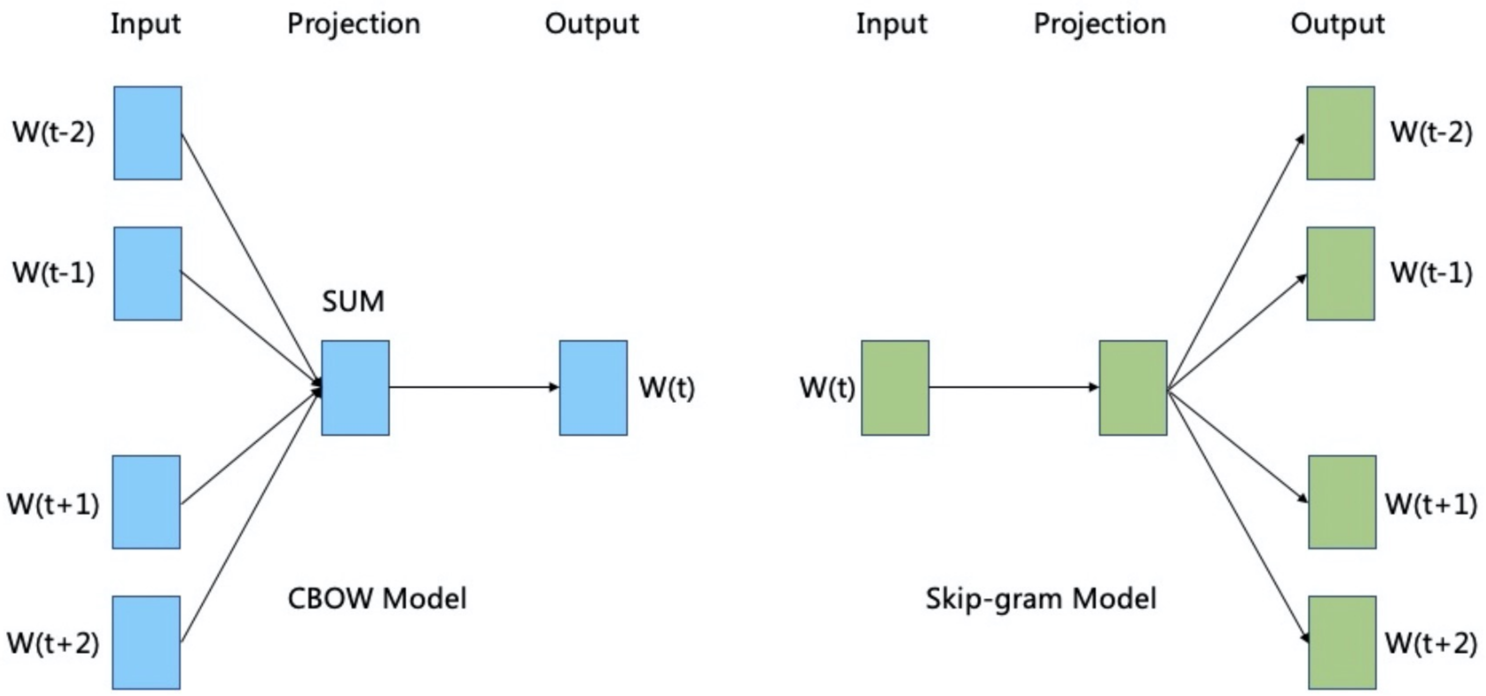
Description Embedding

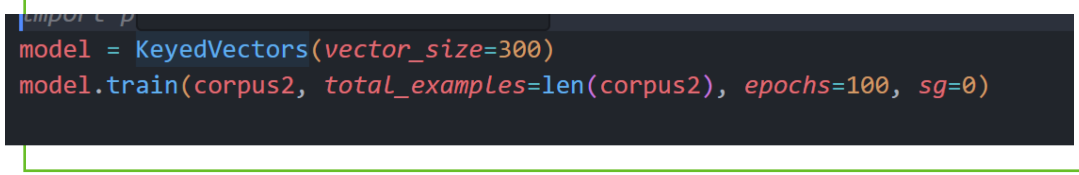
We Train Word2Vec Model Using The Description from all books.

# Word2Vec Based Book Similarity

**Training Word2Vec model**

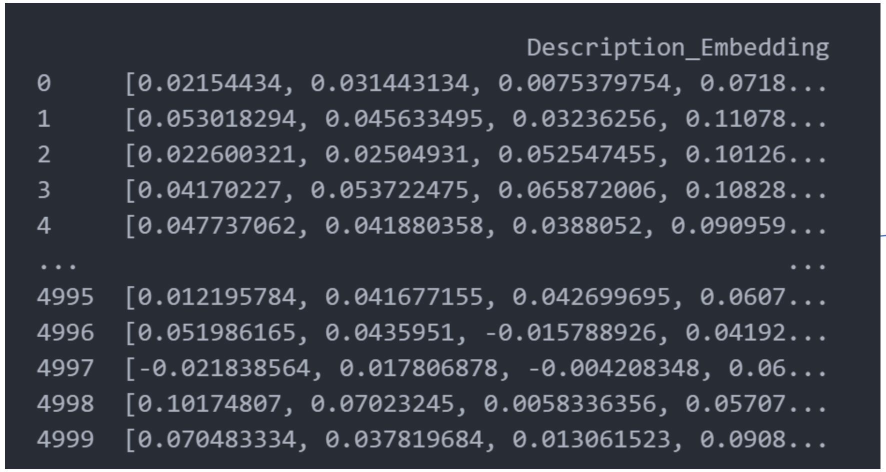
- **We Use CBOW**
- **Train data: All Description from all books**
- **Train 100 epoch**





```
import p
model = KeyedVectors(vector_size=300)
model.train(corpus2, total_examples=len(corpus2), epochs=100, sg=0)
```

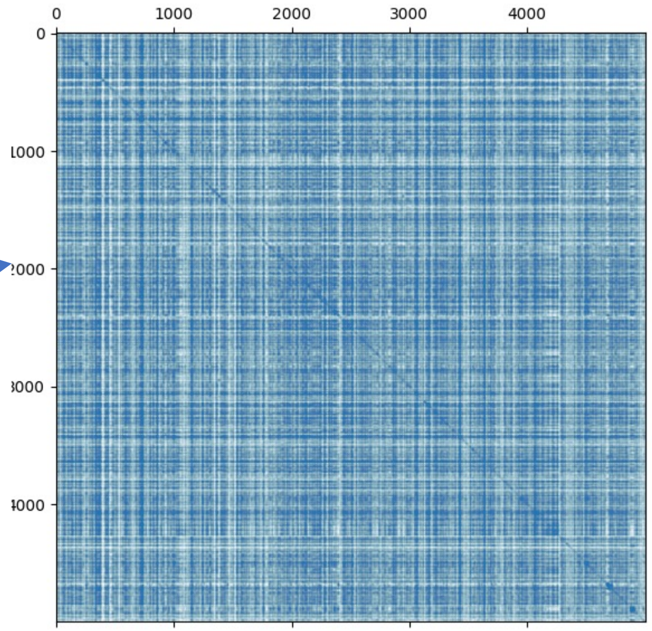
# Word2Vec Based Book Similarity



```
                          Description_Embedding
0     [0.02154434, 0.031443134, 0.0075379754, 0.0718...
1     [0.053018294, 0.045633495, 0.03236256, 0.11078...
2     [0.022600321, 0.02504931, 0.052547455, 0.10126...
3     [0.04170227, 0.053722475, 0.065872006, 0.10828...
4     [0.047737062, 0.041880358, 0.0388052, 0.090959...
...                                                 ...
4995  [0.012195784, 0.041677155, 0.042699695, 0.0607...
4996  [0.051986165, 0.0435951, -0.015788926, 0.04192...
4997  [-0.021838564, 0.017806878, -0.004208348, 0.06...
4998  [0.10174807, 0.07023245, 0.0058336356, 0.05707...
4999  [0.070483334, 0.037819684, 0.013061523, 0.0908...
```

Description Embedding

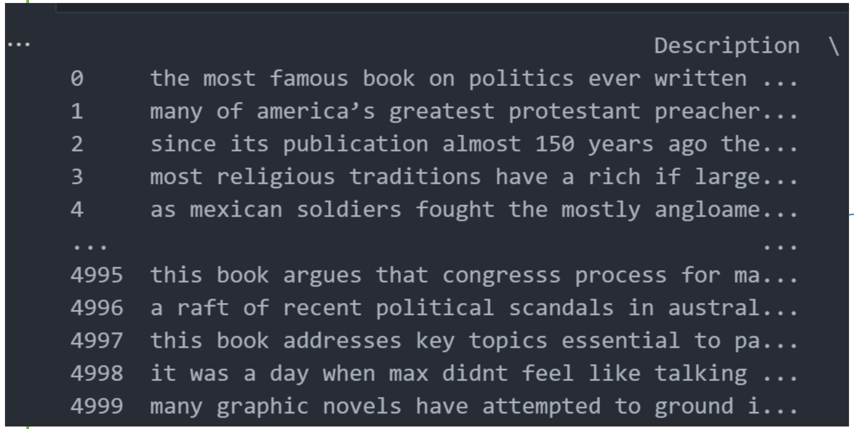
Cosine similarity



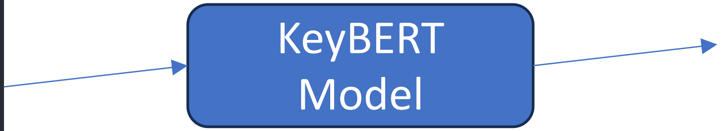
Book Similarity Matrix

CSU | Cleveland State University

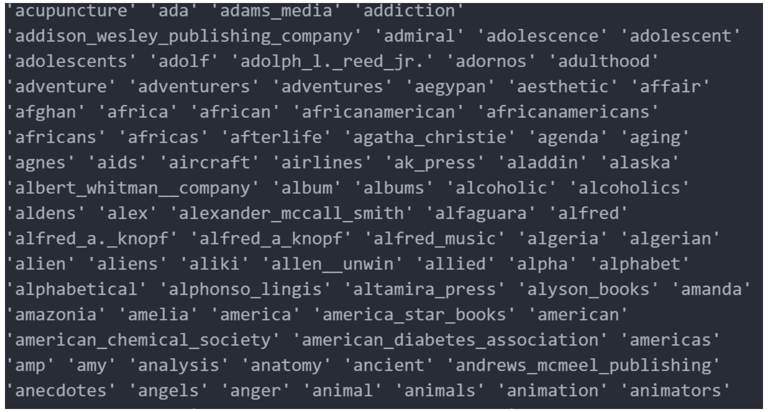
# TF-IDF Based Book Similarity



```
...                                          Description  \
    0     the most famous book on politics ever written ...
    1     many of america’s greatest protestant preacher...
    2     since its publication almost 150 years ago the...
    3     most religious traditions have a rich if large...
    4     as mexican soldiers fought the mostly angloame...
    ...                                                ...
    4995  this book argues that congresss process for ma...
    4996  a raft of recent political scandals in austral...
    4997  this book addresses key topics essential to pa...
    4998  it was a day when max didnt feel like talking ...
    4999  many graphic novels have attempted to ground i...
```



KeyBERT Model



```
'acupuncture' 'ada' 'adams_media' 'addiction'
'addison_wesley_publishing_company' 'admiral' 'adolescence' 'adolescent'
'adolescents' 'adolf' 'adolph_l._reed_jr.' 'adornos' 'adulthood'
'adventure' 'adventurers' 'adventures' 'aegypan' 'aesthetic' 'affair'
'afghan' 'africa' 'african' 'africanamerican' 'africanamericans'
'africans' 'africas' 'afterlife' 'agatha_christie' 'agenda' 'aging'
'agnes' 'aids' 'aircraft' 'airlines' 'ak_press' 'aladdin' 'alaska'
'albert_whitman__company' 'album' 'albums' 'alcoholic' 'alcoholics'
'aldens' 'alex' 'alexander_mccall_smith' 'alfaguara' 'alfred'
'alfred_a._knopf' 'alfred_a_knopf' 'alfred_music' 'algeria' 'algerian'
'alien' 'aliens' 'aliki' 'allen__unwin' 'allied' 'alpha' 'alphabet'
'alphabetical' 'alphonso_lingis' 'altamira_press' 'alyson_books' 'amanda'
'amazonia' 'amelia' 'america' 'america_star_books' 'american'
'american_chemical_society' 'american_diabetes_association' 'americas'
'amp' 'amy' 'analysis' 'anatomy' 'ancient' 'andrews_mcmeel_publishing'
'anecdotes' 'angels' 'anger' 'animal' 'animals' 'animation' 'animators'
```

Processed Book Data Description

We Use KeyBERT Pretrained Model

KeyWords

KeyBERT is a keyword extraction model that leverages the BERT (Bidirectional Encoder Representations from Transformers) architecture to identify important keywords and phrases from a given text.

CSU Cleveland State University

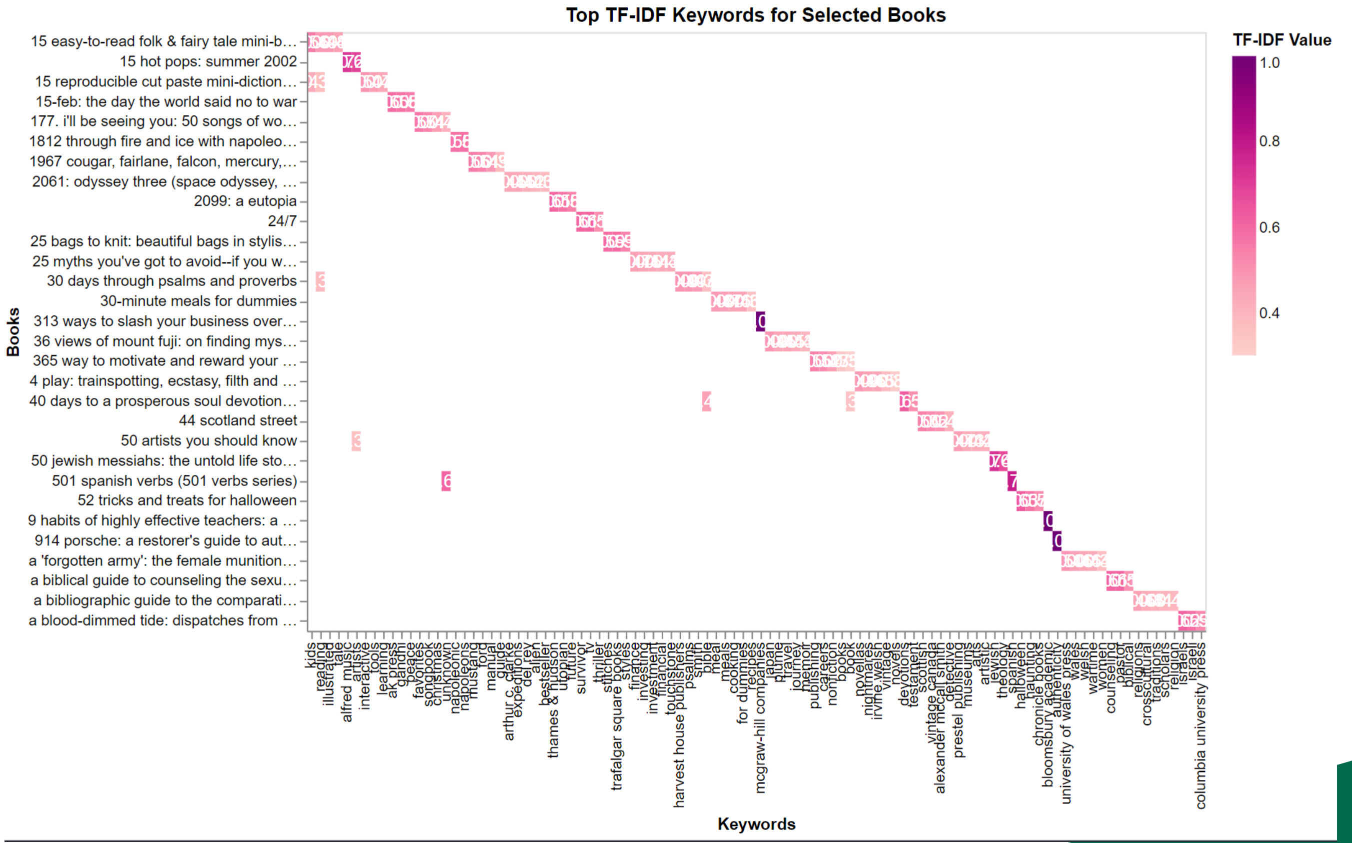
# TF-IDF Based Book Similarity

The top TF-IDF keywords from the selected Books

- The X-axis represents all the keywords.
- The Y-axis are selected books.

**We can find some book have the same keywords. Such as**

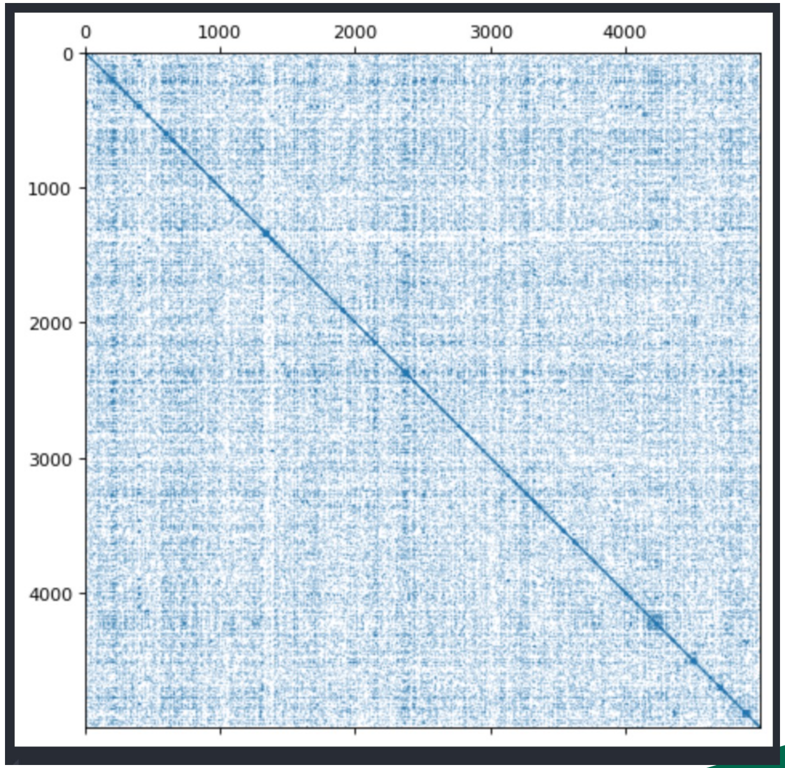
**Book1: 15 easy to read folk & fairy tale mini-b**
**Book2: 15 reproducible cut paste mini-diction**



# TF-IDF Based Book Similarity

Calculate the Similarity From the TF-IDF Encoding

1. Every row and col means the book

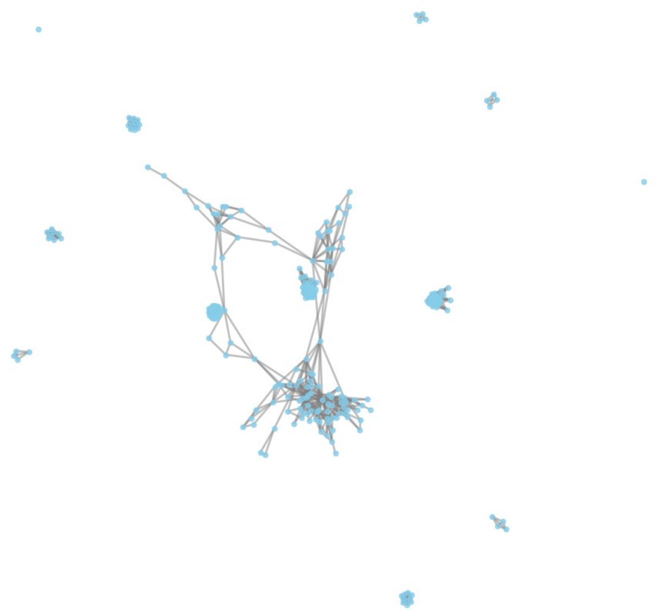


# Generating recommendations

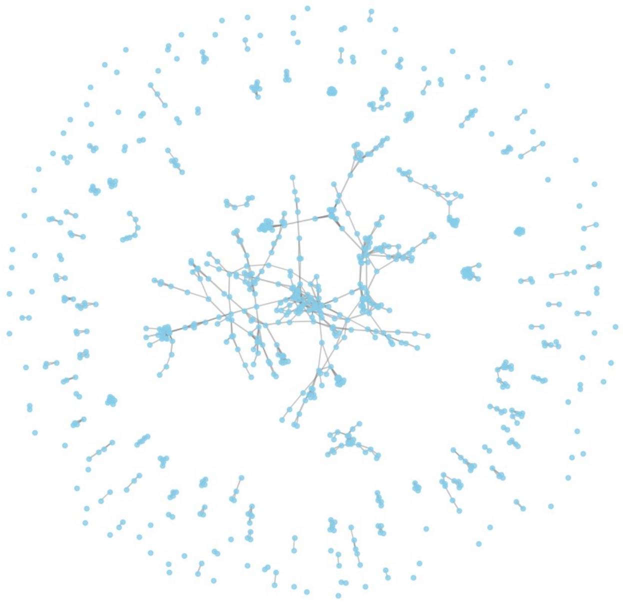
**We have the two Similarity Matrix**

1. Word2Vec Based Similarity Matrix
2. TF-IDF Based Similarity Matrix

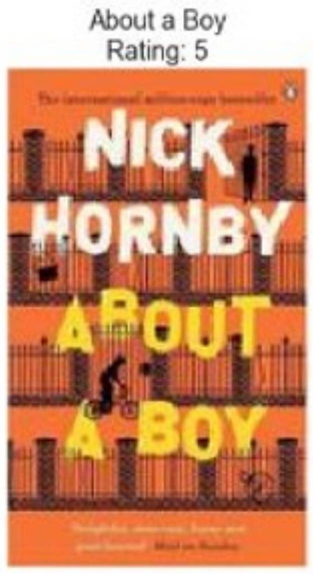
Sparse Book Similarity Network (Higher Threshold) By Word2Vec



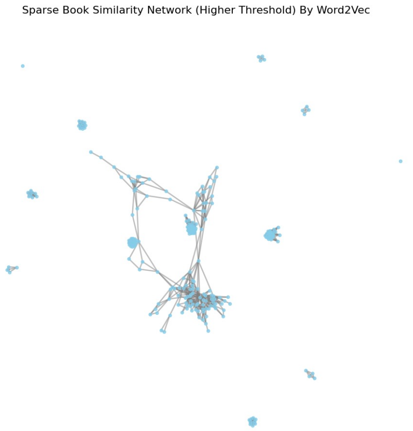
Sparse Book Similarity Network (Higher Threshold) By TF-IDF



# Generating recommendations



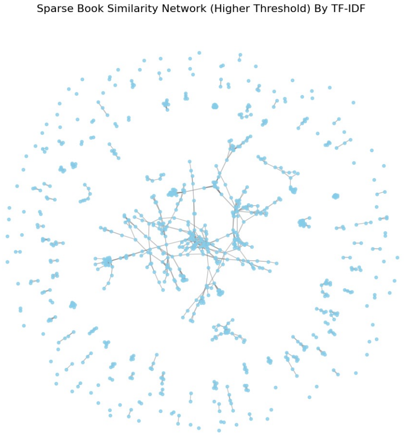
Give a Book


Sparse Book Similarity Network (Higher Threshold) By Word2Vec

Word2Vec Based Similarity

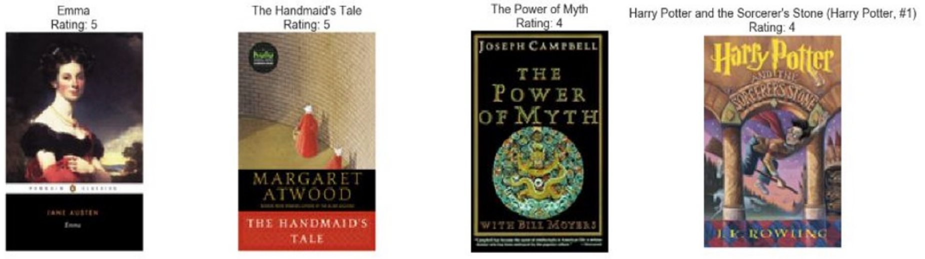
TOP-K Books

The Poisonwood Bible
Rating: 4
The Omnivore's Dilemma: A Natural History of Four Meals
Rating: 4
The Awakening
Rating: 4
Harry Potter and the Goblet of Fire (Harry Potter, #4)
Rating: 4



TF-IDF Based Similarity

TOP-K Books



**If a book appears in both TOP K lists, it should be prioritized for recommendation**

CSU | Cleveland State University

# Book Cover Generation

CSU | Cleveland State University

# Book Cover Benefits

- Beautiful Image Attract Readers
- Highlighting Your Book’s Strengths

# Stable Diffusion

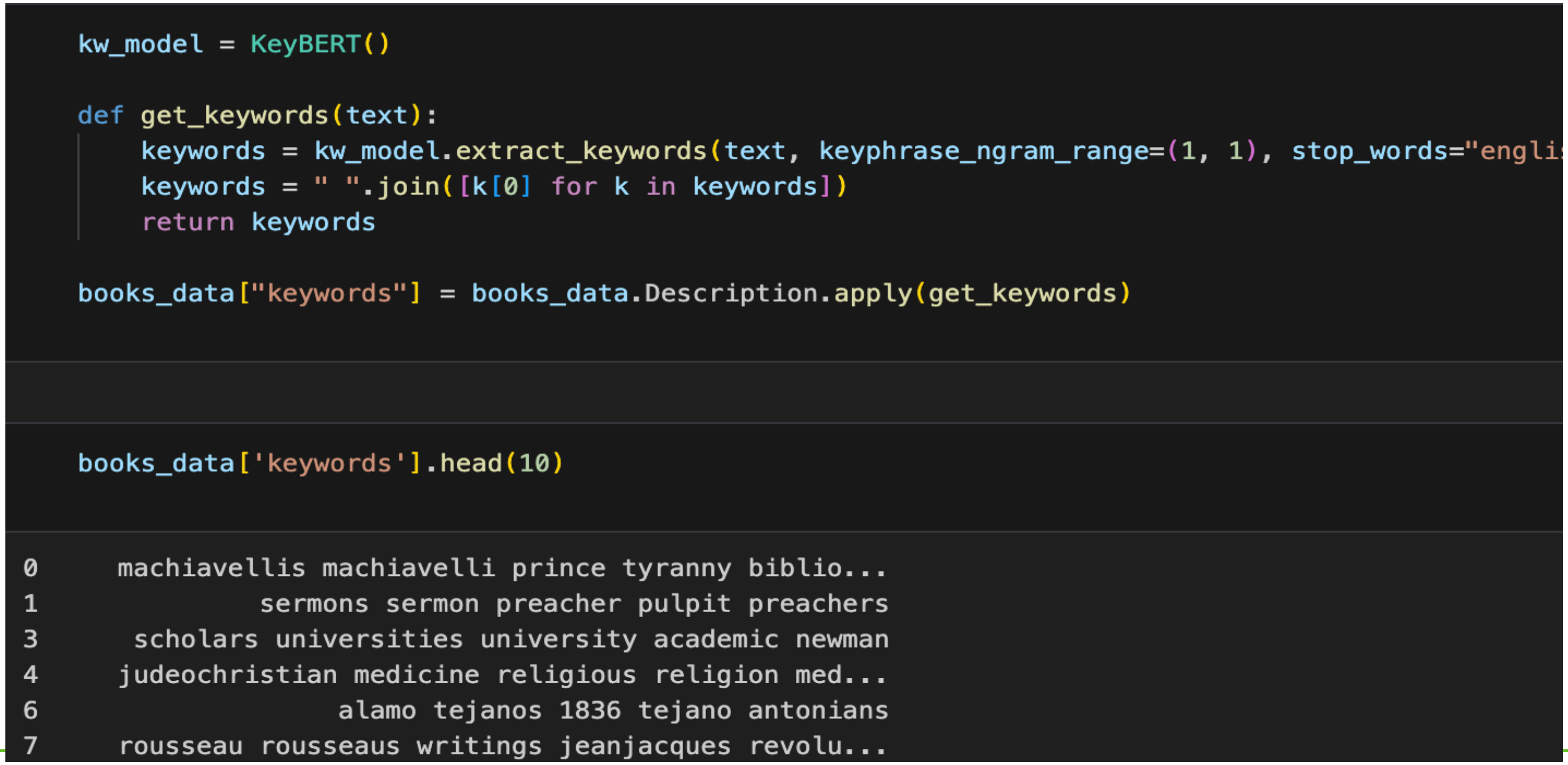
- Understanding human language
- Open source
- Fast

# Problems

- Stable Diffusion Prompt only 75 words
- Book Description more than 75 words

# Extract keywords

Use keyBert to extract keywords for the description of book



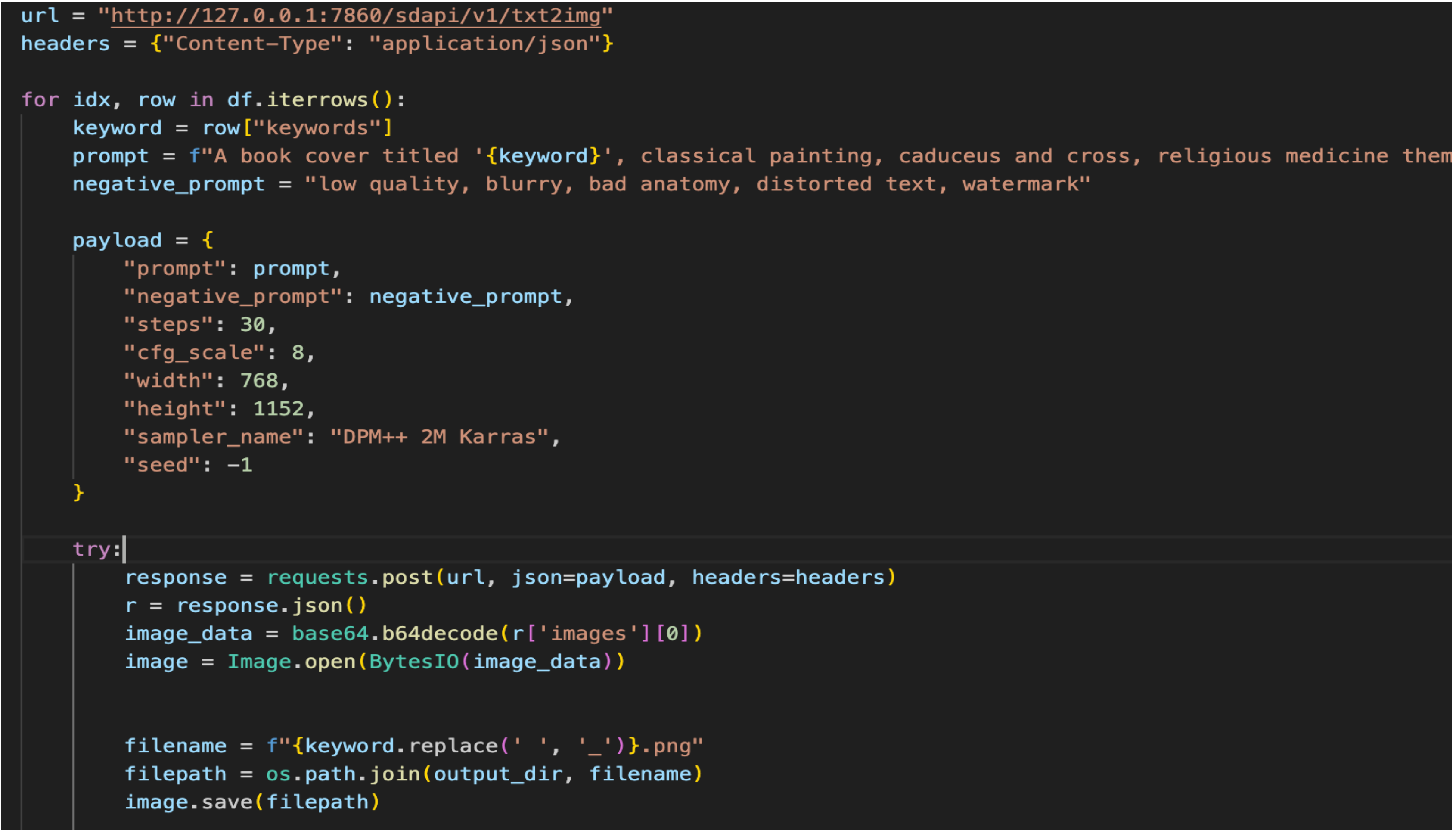
```
kw_model = KeyBERT()

def get_keywords(text):
    keywords = kw_model.extract_keywords(text, keyphrase_ngram_range=(1, 1), stop_words="engli
    keywords = " ".join([k[0] for k in keywords])
    return keywords

books_data["keywords"] = books_data.Description.apply(get_keywords)
```

```
books_data['keywords'].head(10)
```

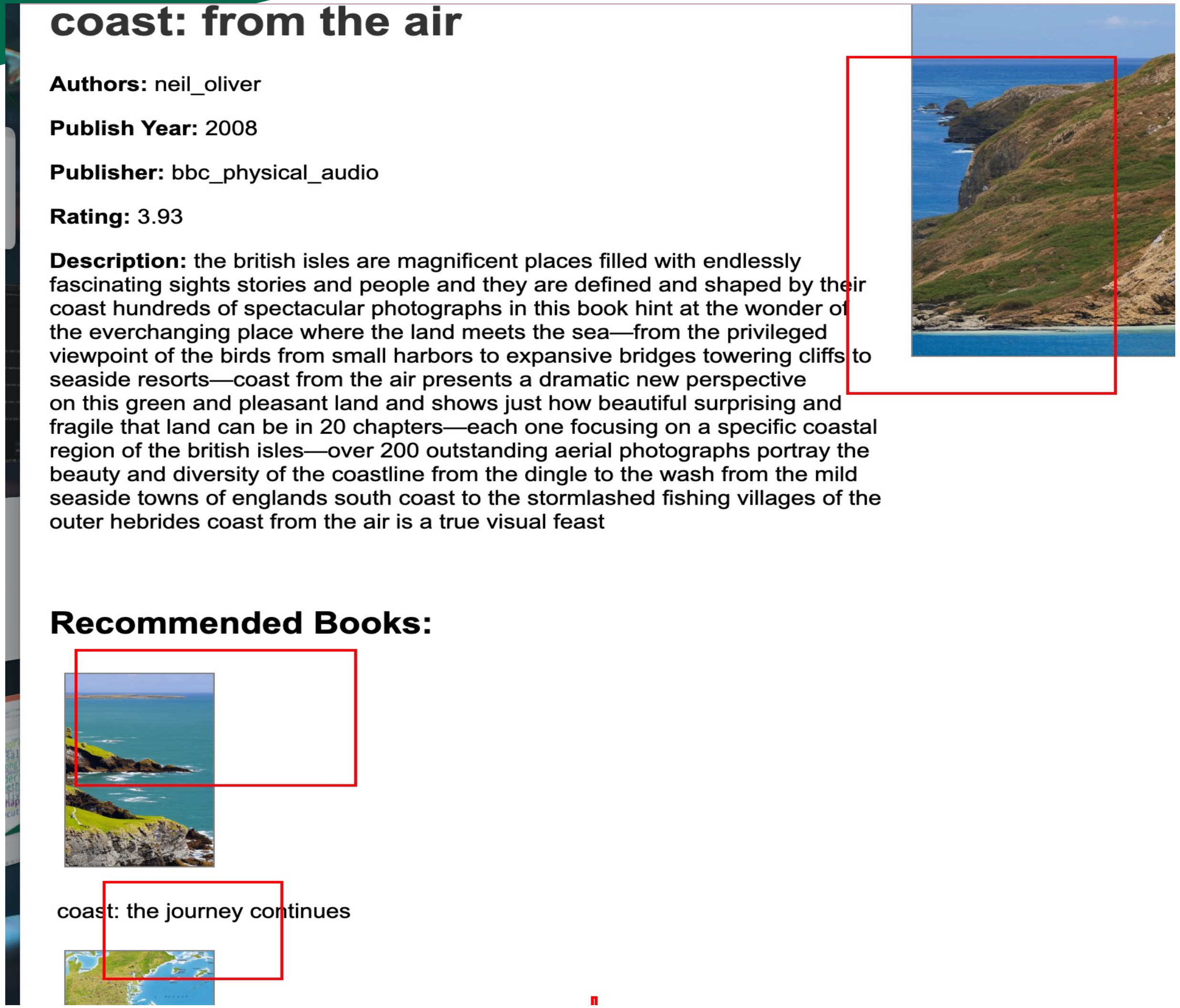
```
0    machiavellis machiavelli prince tyranny biblio...
1             sermons sermon preacher pulpit preachers
3     scholars universities university academic newman
4    judeochristian medicine religious religion med...
6                  alamo tejanos 1836 tejano antonians
7    rousseau rousseaus writings jeanjacques revolu...
```

CSU Cleveland State University

# Generate Book Cover



```
url = "http://127.0.0.1:7860/sdapi/v1/txt2img"
headers = {"Content-Type": "application/json"}

for idx, row in df.iterrows():
    keyword = row["keywords"]
    prompt = f"A book cover titled '{keyword}', classical painting, caduceus and cross, religious medicine them
    negative_prompt = "low quality, blurry, bad anatomy, distorted text, watermark"

    payload = {
        "prompt": prompt,
        "negative_prompt": negative_prompt,
        "steps": 30,
        "cfg_scale": 8,
        "width": 768,
        "height": 1152,
        "sampler_name": "DPM++ 2M Karras",
        "seed": -1
    }

    try:
        response = requests.post(url, json=payload, headers=headers)
        r = response.json()
        image_data = base64.b64decode(r['images'][0])
        image = Image.open(BytesIO(image_data))


        filename = f"{keyword.replace(' ', '_')}.png"
        filepath = os.path.join(output_dir, filename)
        image.save(filepath)
```



# coast: from the air

**Authors:** neil_oliver

**Publish Year:** 2008

**Publisher:** bbc_physical_audio

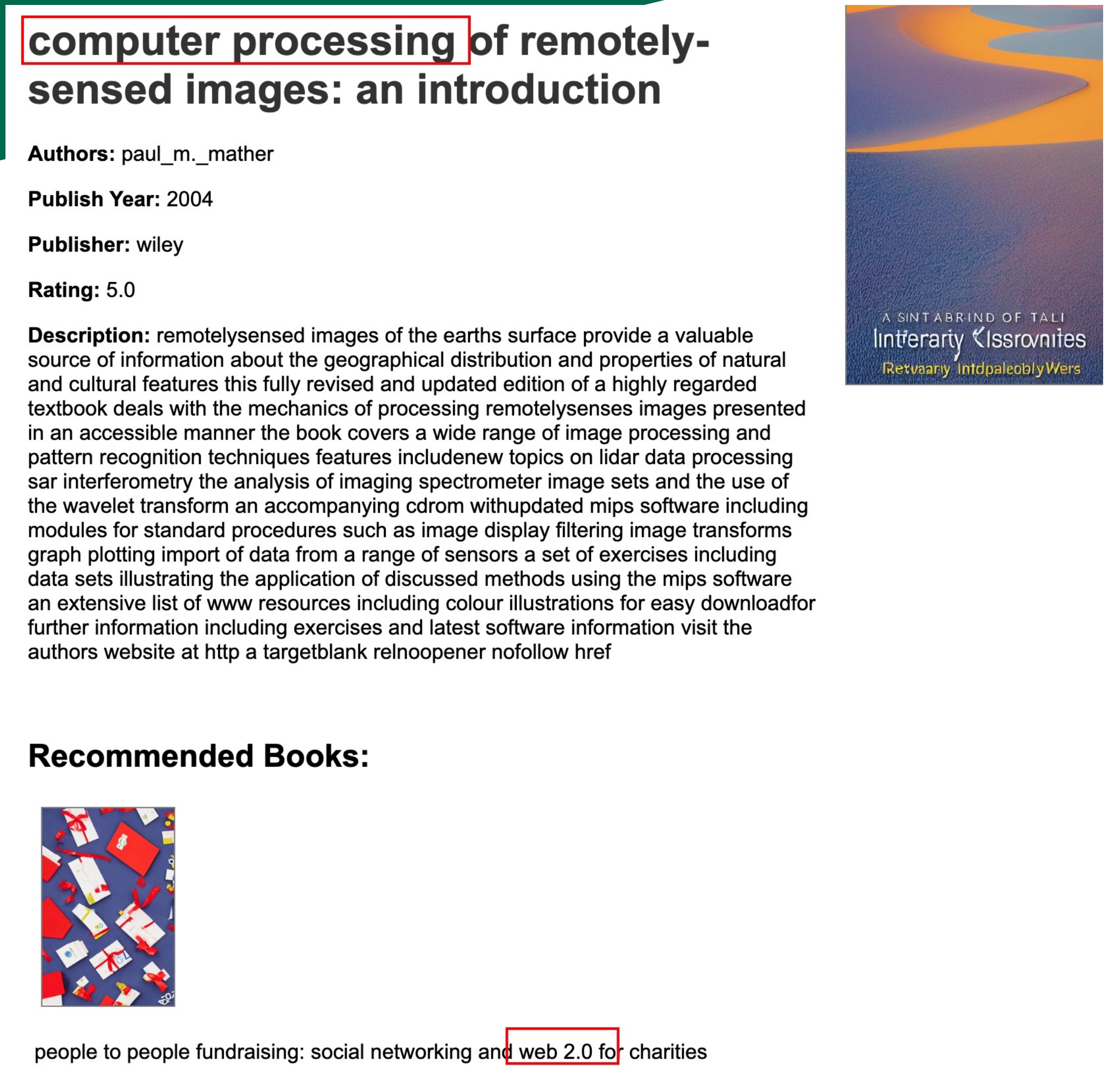
**Rating:** 3.93

**Description:** the british isles are magnificent places filled with endlessly fascinating sights stories and people and they are defined and shaped by their coast hundreds of spectacular photographs in this book hint at the wonder of the everchanging place where the land meets the sea—from the privileged viewpoint of the birds from small harbors to expansive bridges towering cliffs to seaside resorts—coast from the air presents a dramatic new perspective on this green and pleasant land and shows just how beautiful surprising and fragile that land can be in 20 chapters—each one focusing on a specific coastal region of the british isles—over 200 outstanding aerial photographs portray the beauty and diversity of the coastline from the dingle to the wash from the mild seaside towns of englands south coast to the stormlashed fishing villages of the outer hebrides coast from the air is a true visual feast

## Recommended Books:

coast: the journey continues

# Book Recommends



## computer processing of remotely-sensed images: an introduction

**Authors:** paul_m._mather

**Publish Year:** 2004

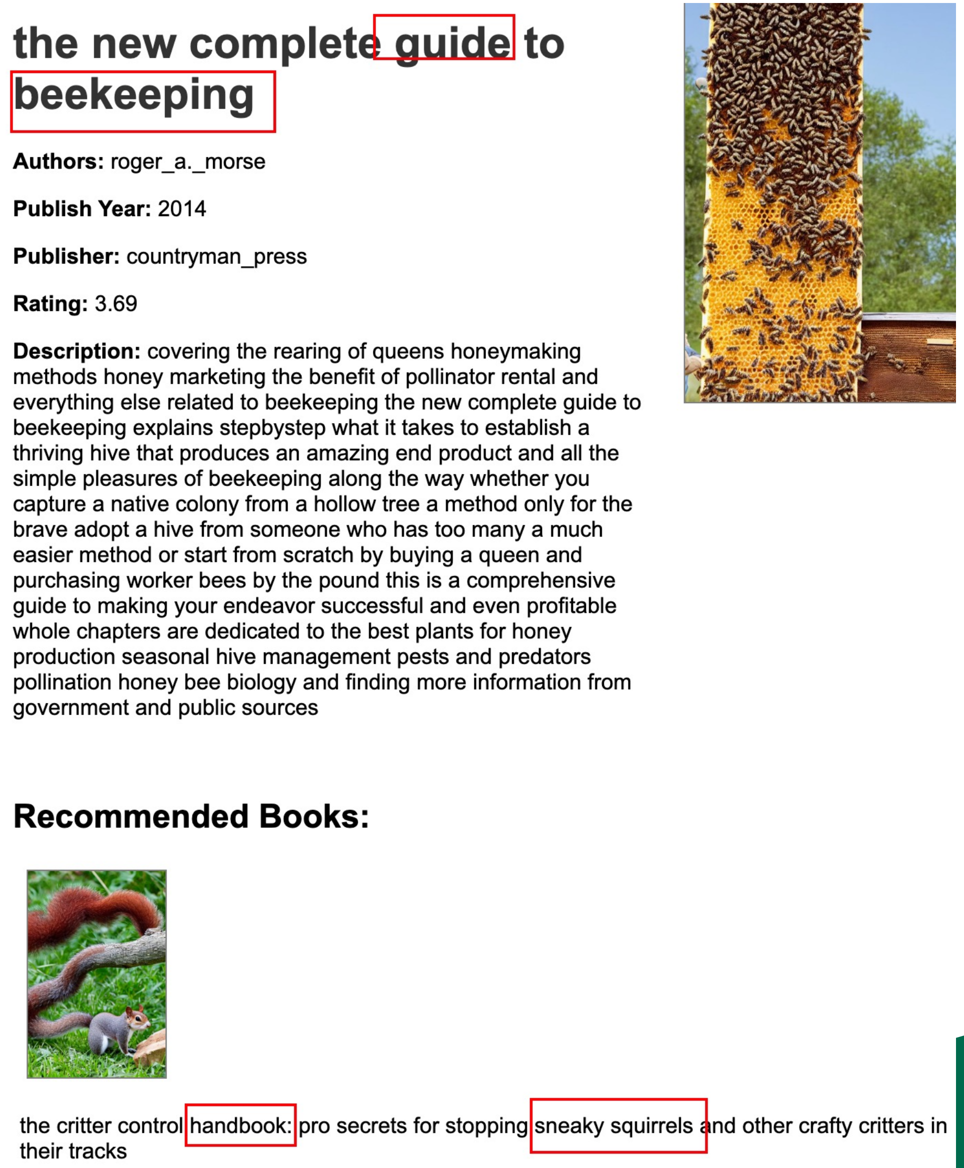
**Publisher:** wiley

**Rating:** 5.0

**Description:** remotelysensed images of the earths surface provide a valuable source of information about the geographical distribution and properties of natural and cultural features this fully revised and updated edition of a highly regarded textbook deals with the mechanics of processing remotelysenses images presented in an accessible manner the book covers a wide range of image processing and pattern recognition techniques features includenew topics on lidar data processing sar interferometry the analysis of imaging spectrometer image sets and the use of the wavelet transform an accompanying cdrom withupdated mips software including modules for standard procedures such as image display filtering image transforms graph plotting import of data from a range of sensors a set of exercises including data sets illustrating the application of discussed methods using the mips software an extensive list of www resources including colour illustrations for easy downloadfor further information including exercises and latest software information visit the authors website at http a targetblank relnoopener nofollow href

### Recommended Books:

people to people fundraising: social networking and web 2.0 for charities



## the new complete guide to beekeeping

**Authors:** roger_a._morse

**Publish Year:** 2014

**Publisher:** countryman_press

**Rating:** 3.69

**Description:** covering the rearing of queens honeymaking methods honey marketing the benefit of pollinator rental and everything else related to beekeeping the new complete guide to beekeeping explains stepbystep what it takes to establish a thriving hive that produces an amazing end product and all the simple pleasures of beekeeping along the way whether you capture a native colony from a hollow tree a method only for the brave adopt a hive from someone who has too many a much easier method or start from scratch by buying a queen and purchasing worker bees by the pound this is a comprehensive guide to making your endeavor successful and even profitable whole chapters are dedicated to the best plants for honey production seasonal hive management pests and predators pollination honey bee biology and finding more information from government and public sources

### Recommended Books:

the critter control handbook: pro secrets for stopping sneaky squirrels and other crafty critters in their tracks

**In this project, we successfully applied Natural Language Processing (NLP) techniques to a book recommendation system, exploring content-based recommendation methods.**

**Summary of Achievements**

- Completed data cleaning and preprocessing.
- Implemented similarity calculations using Word2Vec and TF-IDF.
- Developed a web interface to provide a user-friendly experience.

**Future Work**

- Incorporate More Data: Expand the dataset to enhance the model's generalization ability.
- Optimize Model Algorithms: Experiment with deep learning methods, such as Transformers, to improve text understanding.
- User Behavior Analysis: Integrate user click and review data to refine the recommendation strategy.

# References

- - Goodreads Book Dataset. Retrieved from: **https://www.goodreads.com**
- - Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector **Space. arXiv preprint arXiv:1301.3781.**
- - Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.**
- - Grootendorst, M. (2020). KeyBERT: Minimal and Easy Keyword Extraction with BERT. GitHub repository: **https://github.com/MaartenGr/KeyBERT**
- - Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management.
- - Stable Diffusion: A latent text-to-image diffusion model. Available at: **https://github.com/CompVis/stable-diffusion**