

```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: Books = pd.read_csv('C:/Users/Gaurav/Desktop/Goodreadss_Books.csv')

In [10]: Books.head(1)

Out[10]:
```

Unnamed: 0	bookid	title	author	series	description	genres	awards	characters	places	...	publish_date	num_pages	num_ratings	num_review
0	0	1	Harry Potter and the Half-Blood Prince	J.K. Rowling	Harry Potter #6	The war against Voldemort is not going well; e...	Art,Biography,Business,Children's,Christian,Cl...	Locus Award Nominee for Best Young Adult Novel...	Draco Malfoy,Ron Weasley,Petunia Dursley,Vernon...	Hogwarts School of Witchcraft and Wizardry,Eng...	September 16th 2006	652.0	2553909.0	41470.0

1 rows × 24 columns

```
In [5]: Books.shape

Out[5]: (20068, 24)

In [11]: Books['language'].value_counts()

Out[11]:
```

English	16321
Spanish	449
German	195
French	170
Japanese	47
Multiple languages	24
Italian	14
Chinese	14
Portuguese	13
Greek, Ancient to 1453	10
Russian	9
Latin	4
Arabic	4
Dutch	3
Turkish	2
Danish	2
English, Middle 1100-1500	2
Greek, Modern 1453-	2
Munda languages	1
Norwegian	1
Ladino	1
Serbian	1

Name: language, dtype: int64

```
In [12]: books=Books[['bookId', 'title', 'author', 'description', 'genres', 'language', 'avg_rating']]

In [127]: books.head()

Out[127]:
```

	bookid	title	author	description	genres	language	avg_rating
0	1	Harry Potter and the Half-Blood Prince	J.K. Rowling	The war against Voldemort is not going well; e...	Art,Biography,Business,Children's,Christian,Cl...	English	4.57
1	2	Harry Potter and the Order of the Phoenix	J.K. Rowling,Mary GrandPré	There is a door at the end of a silent corrido...	Art,Biography,Business,Children's,Christian,Cl...	English	4.50
2	3	Harry Potter and the Sorcerer's Stone	J.K. Rowling,Mary GrandPré	Harry Potter's life is miserable. His parents ...	Art,Biography,Business,Children's,Christian,Cl...	English	4.48
3	4	Harry Potter and the Chamber of Secrets	J.K. Rowling	The Dursleys were so mean and hideous that sum...	Art,Biography,Business,Children's,Christian,Cl...	English	4.43
4	5	Harry Potter and the Prisoner of Azkaban	J.K. Rowling,Mary GrandPré	For twelve long years, the dread fortress of A...	Art,Biography,Business,Children's,Christian,Cl...	English	4.57

```
In [13]: books.isnull().sum()

Out[13]:
```

bookId	0
title	559
author	561
description	2687
genres	256
language	2779
avg_rating	559

dtype: int64

```
In [14]: books.dropna(inplace=True)

C:\Users\Gaurav\AppData\Local\Temp\ipykernel_11100\2979460607.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
books.dropna(inplace=True)

In [15]: books.shape

Out[15]: (15931, 7)

In [16]: books.duplicated().sum()

Out[16]: 109

In [17]: books=books.drop_duplicates()

In [18]: books.duplicated().sum()

Out[18]: 0

In [19]: books.head(1)

Out[19]:
```

	bookid	title	author	description	genres	language	avg_rating
0	1	Harry Potter and the Half-Blood Prince	J.K. Rowling	The war against Voldemort is not going well; e...	Art,Biography,Business,Children's,Christian,Cl...	English	4.57

```
In [20]: books.shape

Out[20]: (15822, 7)

In [21]: books['genres']=books['genres'].map(lambda x:str(x).split(','))
books.head()

Out[21]:
```

	bookid	title	author	description	genres	language	avg_rating
0	1	Harry Potter and the Half-Blood Prince	J.K. Rowling	The war against Voldemort is not going well; e...	[Art, Biography, Business, Children's, Christi...	English	4.57
1	2	Harry Potter and the Order of the Phoenix	J.K. Rowling,Mary GrandPré	There is a door at the end of a silent corrido...	[Art, Biography, Business, Children's, Christi...	English	4.50
2	3	Harry Potter and the Sorcerer's Stone	J.K. Rowling,Mary GrandPré	Harry Potter's life is miserable. His parents ...	[Art, Biography, Business, Children's, Christi...	English	4.48
3	4	Harry Potter and the Chamber of Secrets	J.K. Rowling	The Dursleys were so mean and hideous that sum...	[Art, Biography, Business, Children's, Christi...	English	4.43
4	5	Harry Potter and the Prisoner of Azkaban	J.K. Rowling,Mary GrandPré	For twelve long years, the dread fortress of A...	[Art, Biography, Business, Children's, Christi...	English	4.57

```
In [22]: books['author']=books['author'].map(lambda x:str(x).split(','))

In [23]: books['author']

Out[23]:
```

0	[J.K. Rowling]
1	[J.K. Rowling, Mary GrandPré]
2	[J.K. Rowling, Mary GrandPré]
3	[J.K. Rowling]
4	[J.K. Rowling, Mary GrandPré]
...	...
20063	[Dennis Brindell Fradin, Richard Wahl, Len W. ...]
20064	[Delia Sherman, Cortney Skinner, Ellen Kushner]
20065	[Delia Sherman]
20066	[Ellen Kushner, Delia Sherman]
20067	[Delia Sherman, Michael Siefener]

Name: author, Length: 15822, dtype: object

```
In [24]: books['genres']=books['genres'].apply(lambda x:[i.replace(" ","") for i in x])

In [25]: books['author']=books['author'].apply(lambda x:[i.replace(" ","") for i in x])

In [26]: books['avg_rating']=books['avg_rating'].apply(lambda x:str(x).split())

In [27]: books['language']=books['language'].apply(lambda x:str(x).split())

In [28]: books['description']=books['description'].apply(lambda x:str(x).split())

In [29]: books.head(1)

Out[29]:
```

	bookid	title	author	description	genres	language	avg_rating
0	1	Harry Potter and the Half-Blood Prince	[J.K.Rowling]	[The, war, against, Voldemort, is, not, going,...	[Art, Biography, Business, Children's, Christi...	[English]	[4.57]

```
In [30]: books['cdata']=books['author']+books['description']+books['genres']+books['language']+books['avg_rating']

In [31]: n_books=books[['bookId','title','cdata']]

In [32]: n_books

Out[32]:
```

	bookid	title	cdata
0	1	Harry Potter and the Half-Blood Prince	[J.K.Rowling, The, war, against, Voldemort, is...
1	2	Harry Potter and the Order of the Phoenix	[J.K.Rowling, MaryGrandPré, There, is, a, door...
2	3	Harry Potter and the Sorcerer's Stone	[J.K.Rowling, MaryGrandPré, Harry, Potter's, l...
3	4	Harry Potter and the Chamber of Secrets	[J.K.Rowling, The, Dursleys, were, so, mean, a...
4	5	Harry Potter and the Prisoner of Azkaban	[J.K.Rowling, MaryGrandPré, For, twelve, long...
...
20063	19996	Connecticut: In Words and Pictures	[DennisBrindellFradin, RichardWahl, LenW.Meent...
20064	19997	Through a Brazen Mirror	[DeliaSherman, CortneySkinner, EllenKushner, l...
20065	19998	The Porcelain Dove	[DeliaSherman, Narrated, by, the, irrepressibl...
20066	19999	The Fall Of The Kings	[EllenKushner, DeliaSherman, From, School, Lib...
20067	20000	Die Blume der Diener	[DeliaSherman, MichaelSiefener, Der, junge, Wi...

15822 rows × 3 columns

```
In [33]: n_books['cdata']=n_books['cdata'].apply(lambda x:"".join(x))

C:\Users\Gaurav\AppData\Local\Temp\ipykernel_11100\1499995807.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
n_books['cdata']=n_books['cdata'].apply(lambda x:"".join(x))

In [34]: n_books['cdata']

Out[34]:
```

0	J.K.RowlingThewaragainstVoldemortisnotgoingwel...
1	J.K.RowlingMaryGrandPréThereisadoorattheendofa...
2	J.K.RowlingMaryGrandPréHarryPotter'slifeiswise...
3	J.K.RowlingTheDursleysweresomeanandhideousthat...
4	J.K.RowlingMaryGrandPréFortwelvelongyears,thed...
...	...
20063	DennisBrindellFradinRichardWahlLenW.MeentsAbri...
20064	DeliaShermanCortneySkinnerEllenKushnerInamedie...
20065	DeliaShermanNarratedbytheirrepressiblechamberm...
20066	EllenKushnerDeliaShermanFromSchoolLibraryJourn...
20067	DeliaShermanMichaelSiefenerDerJungeWilliamFlow...

Name: cdata, Length: 15822, dtype: object

```
In [35]: n_books['cdata']=n_books['cdata'].apply(lambda x:x.lower())

C:\Users\Gaurav\AppData\Local\Temp\ipykernel_11100\705826575.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
n_books['cdata']=n_books['cdata'].apply(lambda x:x.lower())

In [36]: pip install nltk

Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: nltk in c:\users\gaurav\appdata\roaming\python\python310\site-packages (3.7)
Requirement already satisfied: joblib in c:\users\gaurav\appdata\roaming\python\python310\site-packages (from nltk) (1.1.0)
Requirement already satisfied: tqdm in c:\users\gaurav\appdata\roaming\python\python310\site-packages (from nltk) (4.64.0)
Requirement already satisfied: click in c:\users\gaurav\appdata\roaming\python\python310\site-packages (from nltk) (8.1.3)
Requirement already satisfied: regex>=2021.8.3 in c:\users\gaurav\appdata\roaming\python\python310\site-packages (from nltk) (2022.4.24)
Requirement already satisfied: colorama in c:\users\gaurav\appdata\roaming\python\python310\site-packages (from click->nltk) (0.4.4)
Note: you may need to restart the kernel to use updated packages.

In [37]: import nltk

In [38]: from nltk.stem.porter import PorterStemmer
ps= PorterStemmer()

In [39]: def stem(text):
y=[]

    for i in text.split():
        y.append(ps.stem(i))

    return " ".join(y)

In [40]: n_books['cdata']=n_books['cdata'].apply(stem)

C:\Users\Gaurav\AppData\Local\Temp\ipykernel_11100\295756077.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
n_books['cdata']=n_books['cdata'].apply(stem)

In [41]: import sklearn

In [42]: from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=5000,stop_words='english')

In [44]: cv_matrix=cv.fit_transform(n_books['cdata']).toarray()

In [ ]: cv_matrix

In [ ]: len(cv.get_feature_names())

In [45]: from sklearn.metrics.pairwise import cosine_similarity

In [46]: similarity=cosine_similarity(cv_matrix)

In [51]: sorted(list(enumerate(similarity[0])),reverse=True,key=lambda x:x[1])[1:6]

Out[51]: ((9001, 0.588348405414552),
(14356, 0.588348405414552),
(9464, 0.5547001962252291),
(38941, 0.5393193716300061),
(6313, 0.5017452060042545)]

In [63]: def recommend(books):
    books_index = n_books[n_books['title']==books].index[0]
    distances = similarity[books_index]
    books_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

    for i in books_list:
        print(n_books.iloc[i[0]].title)

In [66]: recommend('Harry Potter and the Half-Blood Prince')

Captivating: Unveiling the Mystery of a Woman's Soul
Romeo and Juliet
The Rebel
The Old Adam: A Story of Adventure
Buffalo Coat

In [62]: n_books.iloc[965].title

Out[62]: "Tyler Florence's Real Kitchen: An Indespensible Guide for Anybody Who Likes to Cook"
```