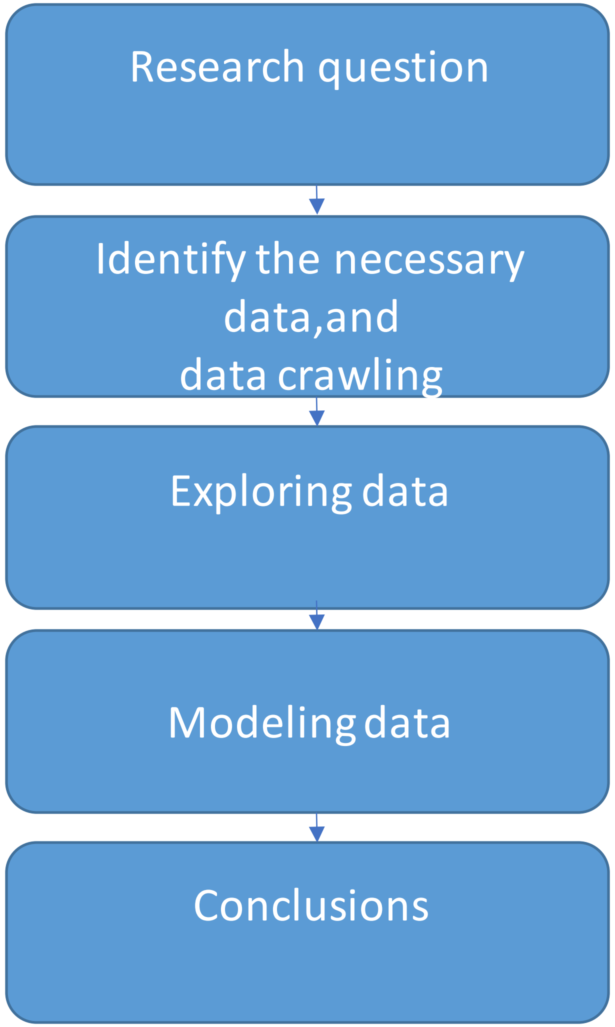
Book Recommendation System

The system will recommend to the user 5 books that he might like based on a book he has already read and loved

Tal hadad

github.com/talhd/Book-Recommendation-System

main steps



Research question

Identify the necessary data,and data crawling

Exploring data

Modeling data

Conclusions

The research question

Is it possible to compare books and quantify the similarity between them?

How can we achieved the most accurate results?

Data crawling

- All the data is taken from goodreads.com the necessary information is:

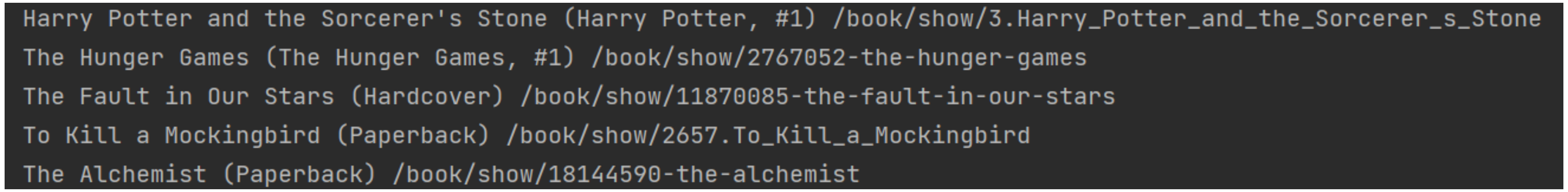
book title,summary,book rating,categories,date of publication

- In the project I used Python packages Beautiful Soup,request

Regular expression- re,and pandas for data crawling

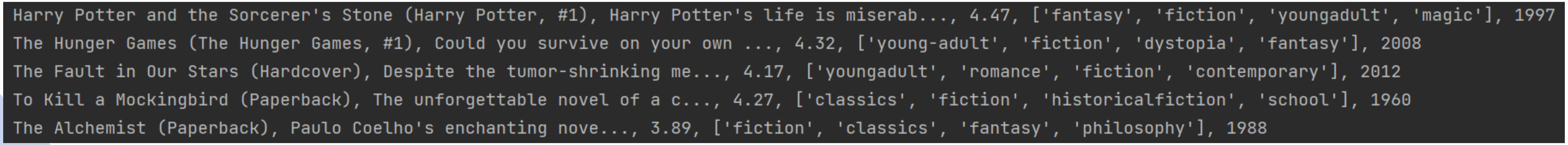
- The process has 2 main stages:

.1Build a list that contain the title of the book and the uri to the book on goodreads.com



```
Harry Potter and the Sorcerer's Stone (Harry Potter, #1) /book/show/3.Harry_Potter_and_the_Sorcerer_s_Stone
The Hunger Games (The Hunger Games, #1) /book/show/2767052-the-hunger-games
The Fault in Our Stars (Hardcover) /book/show/11870085-the-fault-in-our-stars
To Kill a Mockingbird (Paperback) /book/show/2657.To_Kill_a_Mockingbird
The Alchemist (Paperback) /book/show/18144590-the-alchemist
```

.2Gather all the relevant information about the book,for each book on the list



```
Harry Potter and the Sorcerer's Stone (Harry Potter, #1), Harry Potter's life is miserab..., 4.47, ['fantasy', 'fiction', 'youngadult', 'magic'], 1997
The Hunger Games (The Hunger Games, #1), Could you survive on your own ..., 4.32, ['young-adult', 'fiction', 'dystopia', 'fantasy'], 2008
The Fault in Our Stars (Hardcover), Despite the tumor-shrinking me..., 4.17, ['youngadult', 'romance', 'fiction', 'contemporary'], 2012
To Kill a Mockingbird (Paperback), The unforgettable novel of a c..., 4.27, ['classics', 'fiction', 'historicalfiction', 'school'], 1960
The Alchemist (Paperback), Paulo Coelho's enchanting nove..., 3.89, ['fiction', 'classics', 'fantasy', 'philosophy'], 1988
```

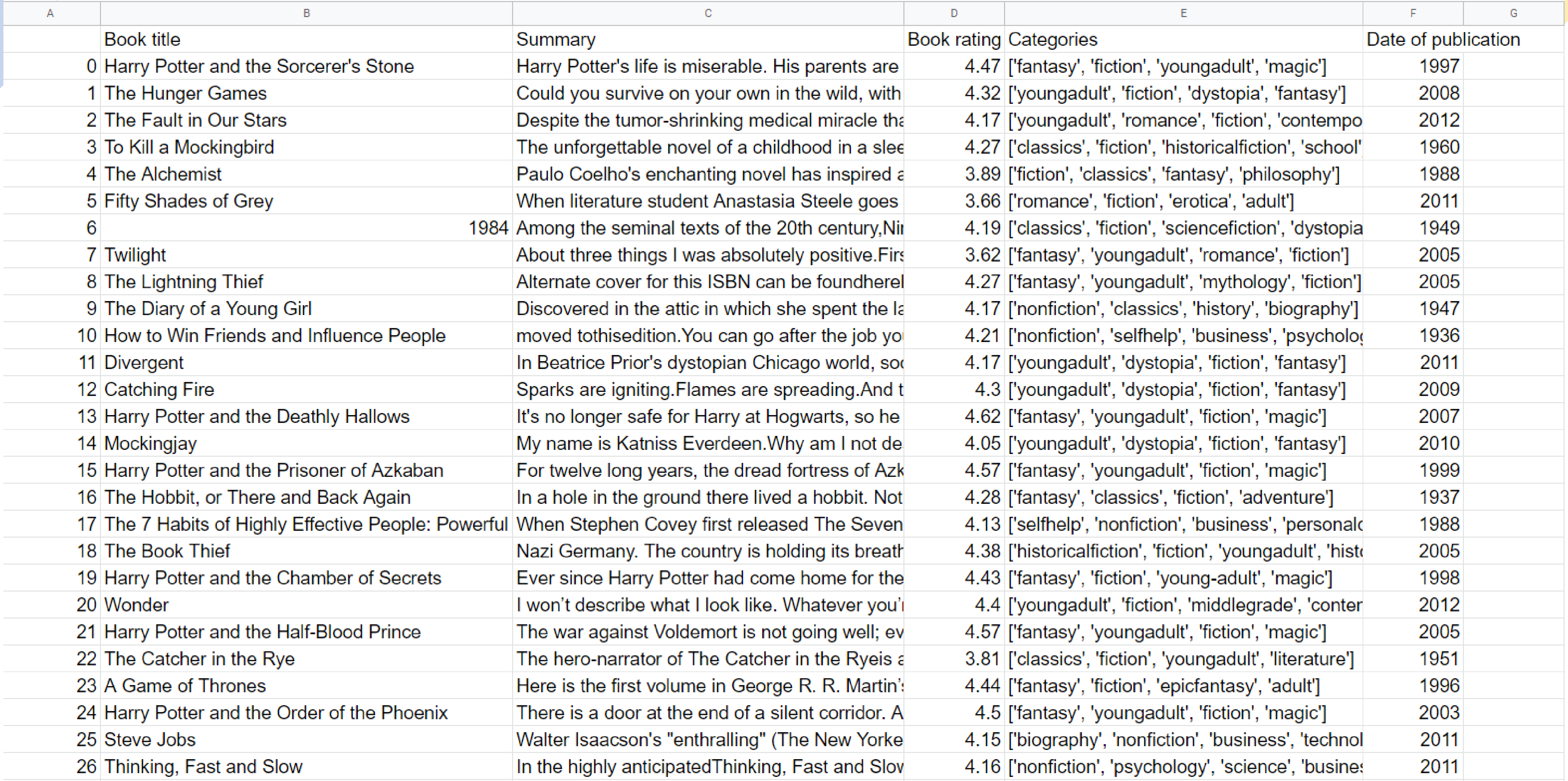
Data cleansing

In the next step we need to perform data cleansing.

- remove data we do not need,like the volume number in the series or comments about the cover
- remove duplication and lines with missing data
- remove books that have not yet been published
- Remov books that are too old (before (1564

Although there is also importance to old books ,we do not want in the system old books to the point of being irrelevant

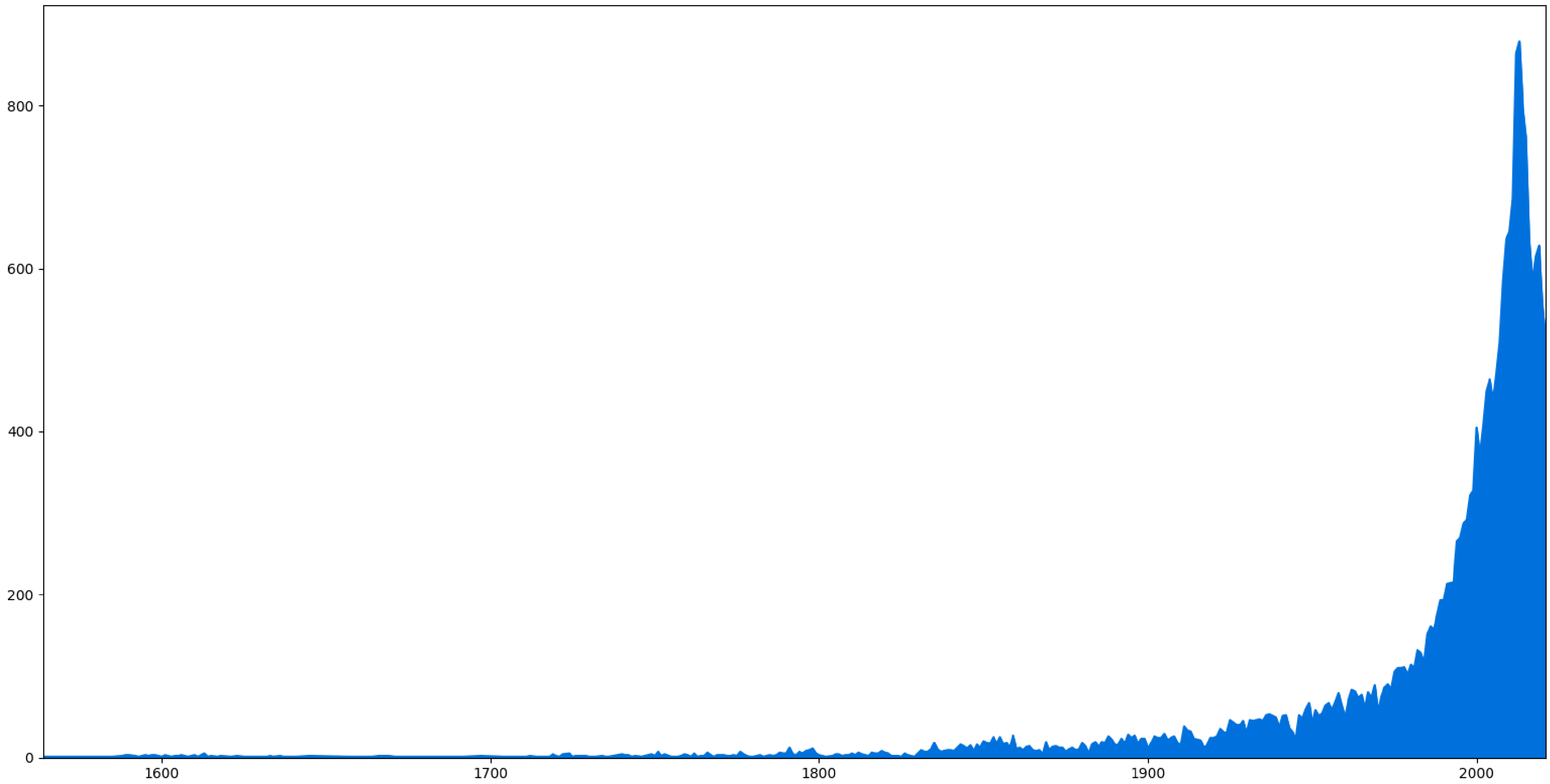
Before data cleansing we had 27,909 books,and then we were left with 22,371

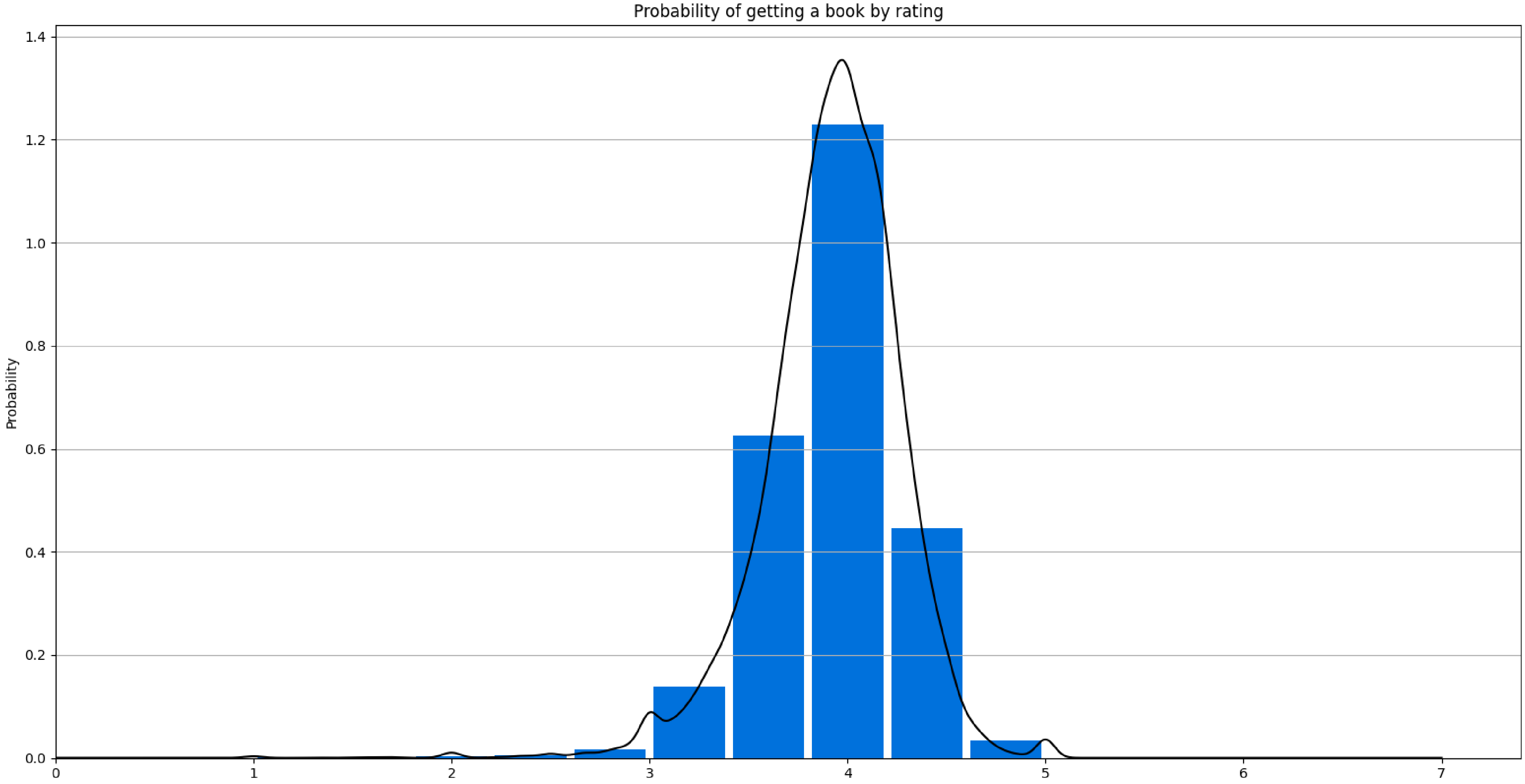


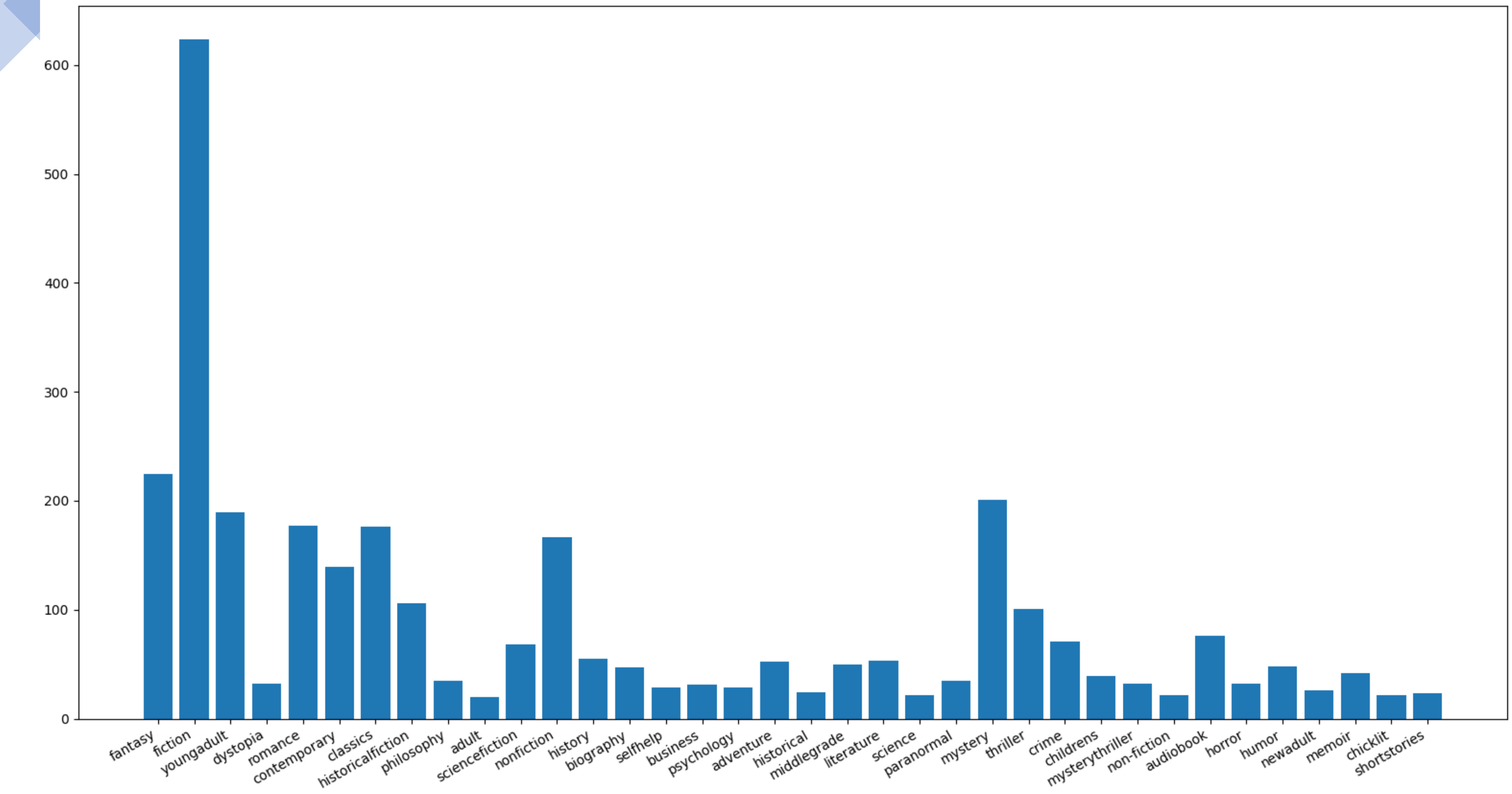
A	B	C	D	E	F	G
	Book title	Summary	Book rating	Categories	Date of publication	
0	Harry Potter and the Sorcerer's Stone	Harry Potter's life is miserable. His parents are	4.47	['fantasy', 'fiction', 'youngadult', 'magic']	1997	
1	The Hunger Games	Could you survive on your own in the wild, with	4.32	['youngadult', 'fiction', 'dystopia', 'fantasy']	2008	
2	The Fault in Our Stars	Despite the tumor-shrinking medical miracle tha	4.17	['youngadult', 'romance', 'fiction', 'contempo	2012	
3	To Kill a Mockingbird	The unforgettable novel of a childhood in a slee	4.27	['classics', 'fiction', 'historicalfiction', 'school'	1960	
4	The Alchemist	Paulo Coelho's enchanting novel has inspired a	3.89	['fiction', 'classics', 'fantasy', 'philosophy']	1988	
5	Fifty Shades of Grey	When literature student Anastasia Steele goes	3.66	['romance', 'fiction', 'erotica', 'adult']	2011	
6	1984	Among the seminal texts of the 20th century,Ni	4.19	['classics', 'fiction', 'sciencefiction', 'dystopia	1949	
7	Twilight	About three things I was absolutely positive.Firs	3.62	['fantasy', 'youngadult', 'romance', 'fiction']	2005	
8	The Lightning Thief	Alternate cover for this ISBN can be foundherel	4.27	['fantasy', 'youngadult', 'mythology', 'fiction']	2005	
9	The Diary of a Young Girl	Discovered in the attic in which she spent the la	4.17	['nonfiction', 'classics', 'history', 'biography']	1947	
10	How to Win Friends and Influence People	moved tothisedition.You can go after the job yo	4.21	['nonfiction', 'selfhelp', 'business', 'psycholo	1936	
11	Divergent	In Beatrice Prior's dystopian Chicago world, soc	4.17	['youngadult', 'dystopia', 'fiction', 'fantasy']	2011	
12	Catching Fire	Sparks are igniting.Flames are spreading.And t	4.3	['youngadult', 'dystopia', 'fiction', 'fantasy']	2009	
13	Harry Potter and the Deathly Hallows	It's no longer safe for Harry at Hogwarts, so he	4.62	['fantasy', 'youngadult', 'fiction', 'magic']	2007	
14	Mockingjay	My name is Katniss Everdeen.Why am I not de	4.05	['youngadult', 'dystopia', 'fiction', 'fantasy']	2010	
15	Harry Potter and the Prisoner of Azkaban	For twelve long years, the dread fortress of Azk	4.57	['fantasy', 'youngadult', 'fiction', 'magic']	1999	
16	The Hobbit, or There and Back Again	In a hole in the ground there lived a hobbit. Not	4.28	['fantasy', 'classics', 'fiction', 'adventure']	1937	
17	The 7 Habits of Highly Effective People: Powerful	When Stephen Covey first released The Seven	4.13	['selfhelp', 'nonfiction', 'business', 'personal	1988	
18	The Book Thief	Nazi Germany. The country is holding its breat	4.38	['historicalfiction', 'fiction', 'youngadult', 'hist	2005	
19	Harry Potter and the Chamber of Secrets	Ever since Harry Potter had come home for the	4.43	['fantasy', 'fiction', 'young-adult', 'magic']	1998	
20	Wonder	I won't describe what I look like. Whatever you'	4.4	['youngadult', 'fiction', 'middlegrade', 'conter	2012	
21	Harry Potter and the Half-Blood Prince	The war against Voldemort is not going well; ev	4.57	['fantasy', 'youngadult', 'fiction', 'magic']	2005	
22	The Catcher in the Rye	The hero-narrator of The Catcher in the Ryeis a	3.81	['classics', 'fiction', 'youngadult', 'literature']	1951	
23	A Game of Thrones	Here is the first volume in George R. R. Martin's	4.44	['fantasy', 'fiction', 'epicfantasy', 'adult']	1996	
24	Harry Potter and the Order of the Phoenix	There is a door at the end of a silent corridor. A	4.5	['fantasy', 'youngadult', 'fiction', 'magic']	2003	
25	Steve Jobs	Walter Isaacson's "enthralling" (The New Yorke	4.15	['biography', 'nonfiction', 'business', 'technol	2011	
26	Thinking, Fast and Slow	In the highly anticipatedThinking, Fast and Slov	4.16	['nonfiction', 'psychology', 'science', 'busine	2011	

The problems that were discovered when we exploring the data

- (1What is the weight given to each feature?
- (2Will we use all the information to achieve a high level of accuracy?
- (3Is there a range of relevant feature at all to arrive to a good match?


800
600
400
200
0
1600
1700
1800
1900
2000


Probability of getting a book by rating
Probability
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5
6
7


600
500
400
300
200
100
0
fantasy
fiction
youngadult
dystopia
romance
contemporary
classics
historicalfiction
philosophy
adult
sciencefiction
nonfiction
history
biography
selfhelp
business
psychology
adventure
historical
middlegrade
literature
science
paranormal
mystery
thriller
crime
childrens
mysterythriller
non-fiction
audiobook
horror
humor
newadult
memoir
chicklit
shortstories

An alternative solution based on the book summary

- we understand that in order to reach a good fit we need to rely on the summary of the book
- When we use the book summary we give weight to the words and not to the technical details,In this way we can quantify and compare books well
- The content of the book is the most important feature in our case

tf–idf and cosine similarity

- in general: we will create the word vectors and calculate the angle between the vectors to find the 5 books closest to the original book
- An example of a vector words for a book Harry Potter and the Sorcerer's Stone
-)Due to technical limitations words that appear in the dictionary but not in the summary will not app ear(

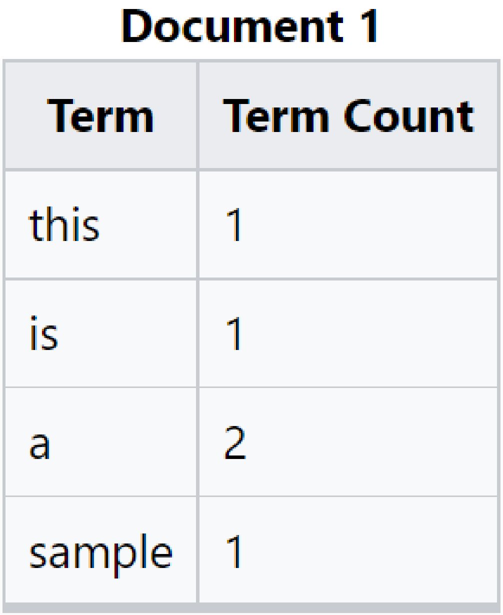
How tf-idf and cosine similarity work

Term Frequency(tf)-the frequency of the word in each text,the ratio of number of times the word appears in a text compared to the total number of words in that text

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$f_{t,d}$ = number of times the word t appears in the text d

for example



Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

$$\mathrm{tf}(''\mathsf{this}'', d_1) = \frac{1}{5} = 0.2$$

-Inverse Data Frequency (idf)
used to calculate the weight of rare words in txts
Rare words get a higher score

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$N = total\ number\ of\ texts$$

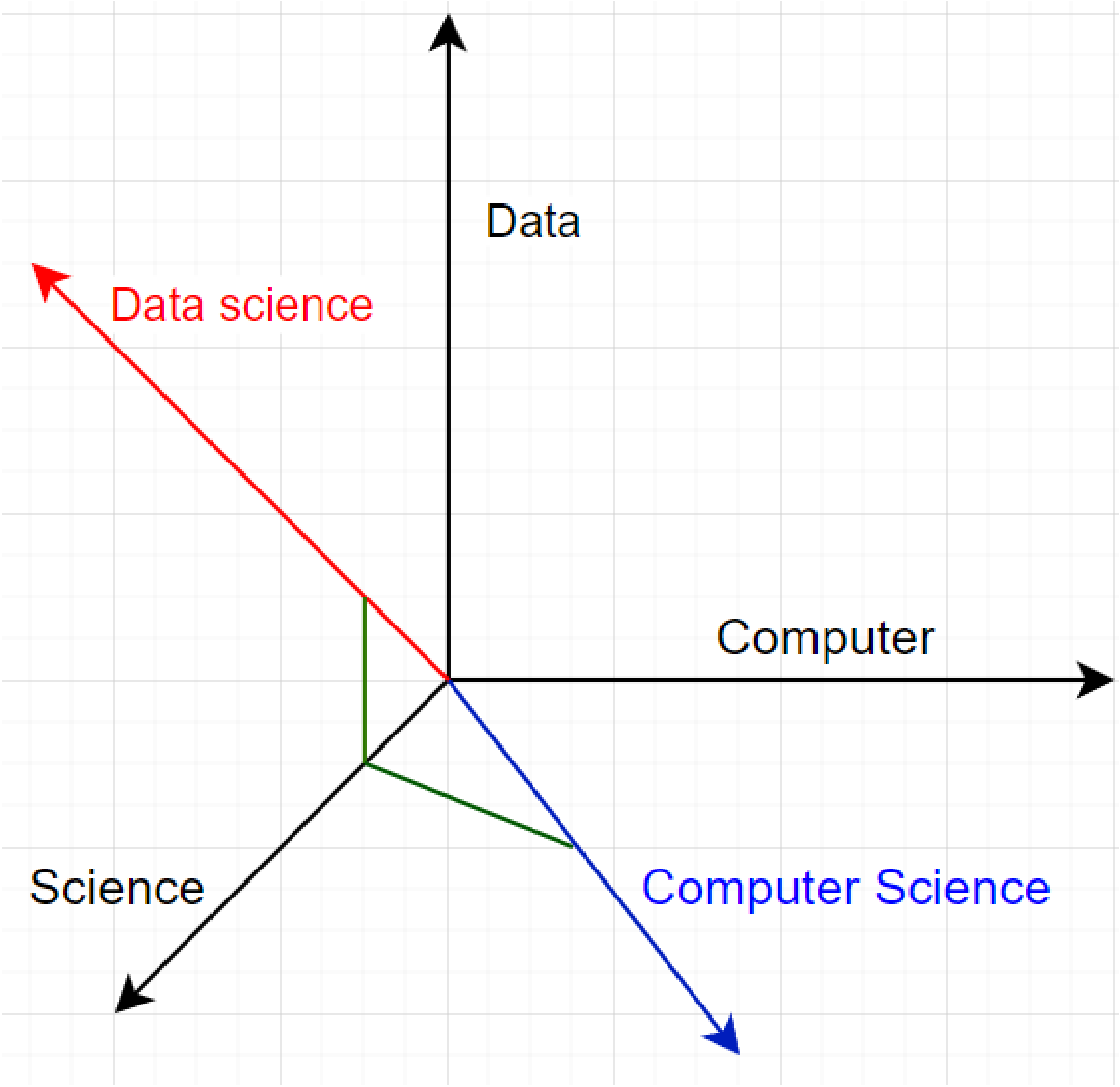
$|\{d \in D : t \in d\}|$ =number of texts in which t appears

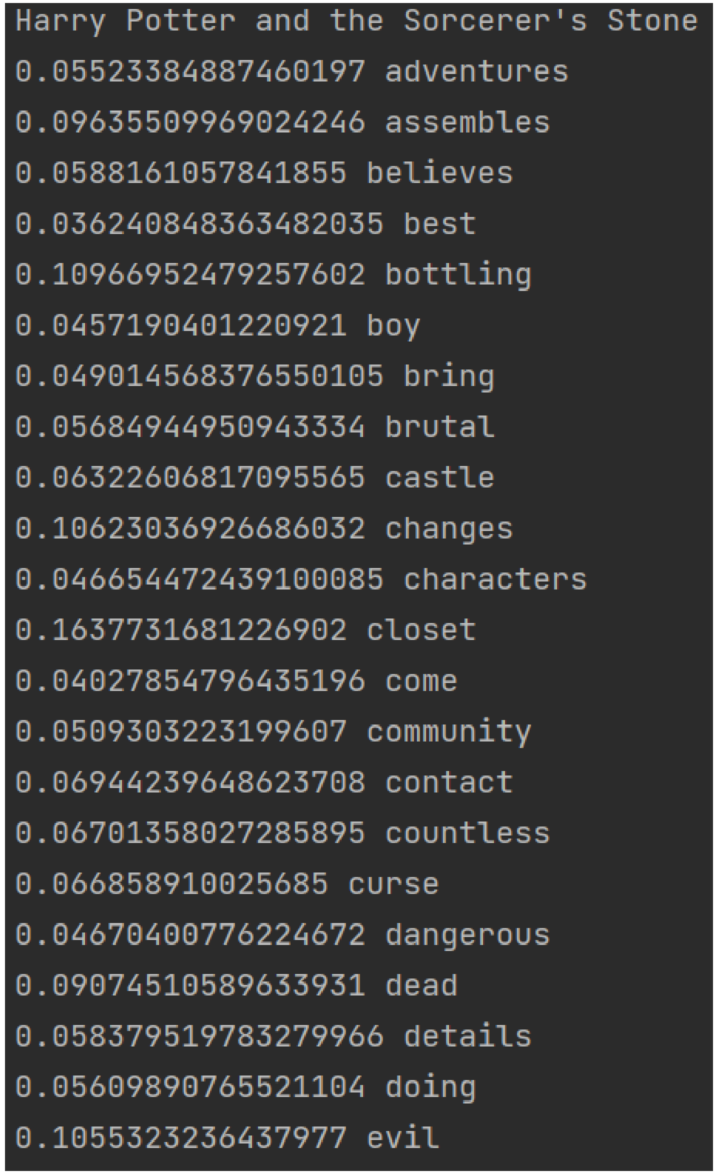
$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Some of the formulas were taken from
https://en.wikipedia.org/wiki/Tf-idf

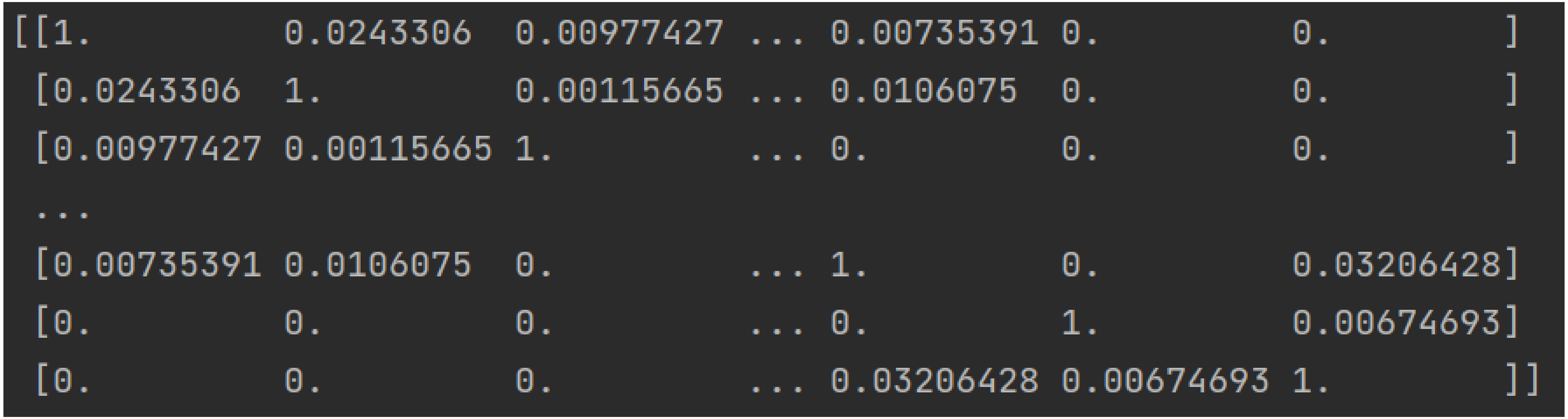
cosine similarity

- Cosine similarity used to calculate the angle between the word vectors,and by doing so we can find the closest vectors to the original vector



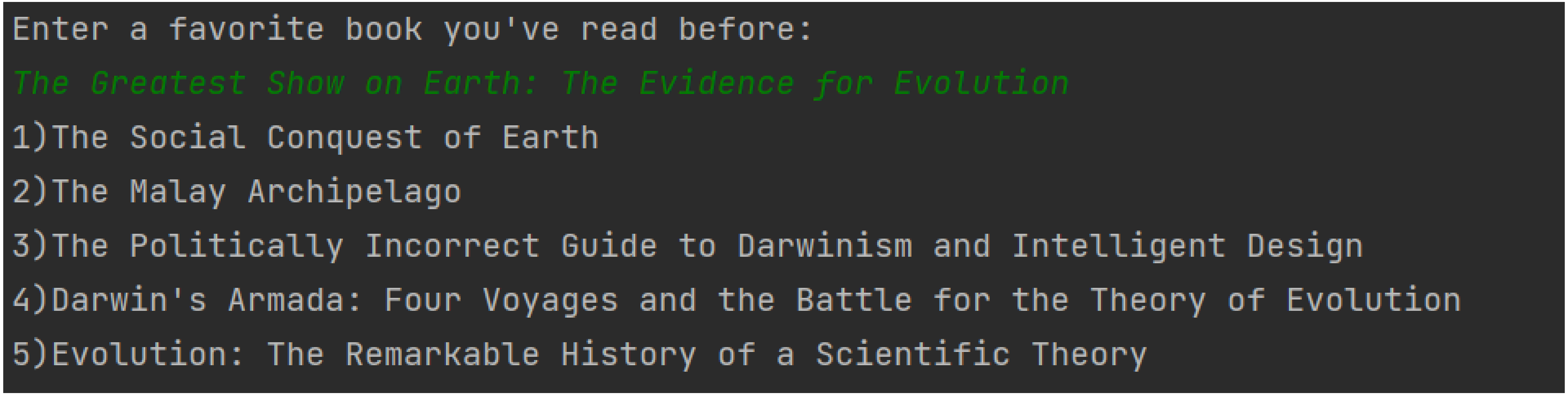


```
Harry Potter and the Sorcerer's Stone
0.05523384887460197 adventures
0.09635509969024246 assembles
0.0588161057841855 believes
0.036240848363482035 best
0.10966952479257602 bottling
0.0457190401220921 boy
0.049014568376550105 bring
0.05684944950943334 brutal
0.06322606817095565 castle
0.10623036926686032 changes
0.046654472439100085 characters
0.1637731681226902 closet
0.04027854796435196 come
0.0509303223199607 community
0.06944239648623708 contact
0.06701358027285895 countless
0.066858910025685 curse
0.04670400776224672 dangerous
0.09074510589633931 dead
0.058379519783279966 details
0.05609890765521104 doing
0.1055323236437977 evil
```



```
[[1.         0.0243306  0.00977427 ... 0.00735391 0.         0.        ]
 [0.0243306  1.         0.00115665 ... 0.0106075  0.         0.        ]
 [0.00977427 0.00115665 1.         ... 0.         0.         0.        ]
 ...
 [0.00735391 0.0106075  0.         ... 1.         0.         0.03206428]
 [0.         0.         0.         ... 0.         1.         0.00674693]
 [0.         0.         0.         ... 0.03206428 0.00674693 1.        ]]
```

Example of using the system



```
Enter a favorite book you've read before:
The Greatest Show on Earth: The Evidence for Evolution
1)The Social Conquest of Earth
2)The Malay Archipelago
3)The Politically Incorrect Guide to Darwinism and Intelligent Design
4)Darwin's Armada: Four Voyages and the Battle for the Theory of Evolution
5)Evolution: The Remarkable History of a Scientific Theory
```

conclusions

You can see that with tf–idf and cosine similarity we solve a lot of problems which appeared when we examined the possibility of using different features to compare books.

In addition to this,in the current form we give weight to the really important data,and ignore data that may influence and lead to wrong result.