

BOOK SEARCH ENGINE

Team Members:

P Sai Sharan
(S20200010174)

P Nachiketa
(S20200010153)

Our Information Retrieval Project Topic was Book Search Engine using some IR Techniques.

The Five Components we applied in our project were mainly:

- a) Indexing
- b) Searching
- C) Refining searches based on specific filters
- d) Capturing FeedBack (Relevance)

Step-1 Importing Necessary Libraries and reading database.csv file.

```
In [2]: import pandas as pd
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\tejad\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!

Out[2]: True
```

```
In [3]: books_data = pd.read_csv('database.csv')
books_data1 = books_data.copy()
books_data1
```

Step-2: Dropping the columns and rows which are not necessary.

```
In [4]: books_data1.drop(['isbn13', 'isbn10', 'thumbnail', 'description', 'published_year', 'ratings_count',
```

Step-3: Normalizing and StopWords Removal.

Normalizing, StopWords Removal and Indexing

```
In [11]: StopWords = set(stopwords.words('english'))

def Normalize(key):
    stringlower = key.lower()
    numberlessstring = re.sub(r'\d+', '', stringlower)
    punclessstring = re.sub(r'^\w\s', '', numberlessstring)
    wspacelessstring = punclessstring.strip()

    if wspacelessstring == '':
        return None

    if wspacelessstring in StopWords:
        return None

    return wspacelessstring

class documentId:
    def __init__(self, documentId):
        self.documentId = documentId
        self.next = None
```

Step-4: Indexing for Categories of the Books.

```

dict1 = {}
StopWords = set(stopwords.words('english'))
books_data3 = books_data2.copy()

for i, des in enumerate(books_data3['categories']):
    tokens = word_tokenize(des)
    for wt in tokens:
        token = Normalize(wt)
        if token == None:
            continue
        else:
            wt = token

            if wt in dict1 and dict1[wt][1].documentId < i:
                dict1[wt][1].next = documentId(i)
                dict1[wt][1] = dict1[wt][1].next
            elif wt not in dict1:
                node = documentId(i)
                dict1[wt] = [node, node]

def pii(dict1):
    for key in dict1:
        print(key, end=' ')
        currentnode = dict1[key][0]
        while currentnode != None:
            print(currentnode.documentId, end=' ')
            currentnode = currentnode.next

```

OUTPUT FOR INDEXING BY CATEGORY:

```

fiction 0 2 3 11 12 15 19 20 22 28 30 35 37 50 54 60 64 66 69 72 81 93 95 96 97 98 100 111 1
12 113 114 115 118 119 120 122 123 125 126 128 129 130 131 132 134 135 136 137 141 143 144 1
46 147 148 149 150 151 152 153 155 156 159 160 162 163 164 165 166 168 171 173 174 175 176 1
77 179 180 182 183 184 186 187 190 191 193 196 197 200 201 202 203 204 205 208 209 211 212 2
14 215 216 217 228 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 251 252 2
53 255 257 258 259 260 261 262 263 264 265 267 268 269 270 271 272 273 274 277 278 279 281 2
82 283 285 288 291 294 295 297 298 300 301 302 303 304 305 307 308 310 311 312 313 314 315 3
17 319 322 328 329 333 335 337 341 343 344 347 348 349 350 351 352 353 355 357 358 359 360 3
65 369 372 374 377 378 380 383 384 385 386 387 388 389 390 391 393 395 397 398 399 400 401 4
02 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 4
25 428 430 432 433 435 436 437 438 439 441 442 443 444 445 449 454 455 457 458 459 461 467 4
68 473 474 475 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 4
96 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 514 515 516 518 521 538 540 5
41 542 545 546 554 557 566 576 590 596 597 601 603 609 611 614 617 620 622 626 640 641 642 6
43 645 647 649 652 653 655 658 663 664 665 666 673 675 676 677 678 679 684 685 688 689 690 6
91 693 695 703 705 706 708 709 710 711 715 718 719 721 722 726 729 730 731 732 735 736 738 7
39 740 741 744 745 746 747 750 757 758 759 760 761 764 768 769 770 771 774 775 776 778 779 7
80 782 783 784 786 787 796 797 798 799 800 801 803 810 811 812 815 819 830 831 832 833 836 8
39 840 842 843 845 846 850 851 852 853 854 856 860 866 867 874 878 885 886 887 888 889 890 8

```

OUTPUT FOR SEARCH BY CATEGORY:

Retrieves all the Books related to the category
drama

```

average_rating      4.3
num_pages           608.0
id                  3594
Name: 3594, dtype: object

title               Witness for the Prosecution & Selected Plays
subtitle            not found
authors             Agatha Christie
categories           English drama
average_rating      4.05
num_pages           352.0
id                  17
Name: 17, dtype: object

title               Fat Pig
subtitle            A Play
authors             Neil LaBute
categories           Drama
average_rating      3.56
num_pages           112.0

```

OUTPUT FOR INDEXING BY TITLE:

```

gilead 0
spider 1 114 432 1522 4438 6068
web 1 306 307 425 501 1272 1681 2786 5200 5662 5666 6364
one 2 168 364 419 680 734 890 899 910 929 1117 1165 1474 1540 1596 1665 1686 1720 1840 1951
2091 2571 2844 2904 3316 3699 4147 4412 4706 4831 6153 6270 6298 6605 6689 6732
tree 2 33 39 418 533 2963 4176 4927 5049 5842
rage 3 5372
angels 3 652 1038 3023 3681 3964 5898 6114 6166 6377
four 4 48 314 500 1104 1185 1460 1484 1609 1712 2783 2795 3073 3259 4821 4969 5295 5354 6766
loves 4 1139 1515 3605
problem 5 1806
pain 5
autobiography 6 371 915 1381 1472 2906 3071 3212
empires 7
monsoon 7 1725
gap 8
madness 8 702 3136 3753 5370 5508 5628
master 9 1213 2167 2614 2623 2633 2649 3648 4346 5109 5863
game 9 181 1495 1718 2416 3241 4072 4238 6737

```

OUTPUT FOR SEARCHING BY TITLE:

Retrieves all the Books related to the query as
Wings of Fire(Title of a Book)

```

////////////////////////////////////
wings of fire
////////////////////////////////////
wings 1606 1694 2158 3010 5073
fire 549 1306 1436 1606 1738 1830 1973 2437 2595 2793 3032 3426 3455 3484 3874 4100 4464 450
8 5199 5514 5519 5668 6593

{4100, 2437, 5514, 5519, 1306, 3484, 1436, 1694, 4508, 3874, 2595, 5668, 549, 1830, 1973, 65
93, 3010, 1606, 1738, 5199, 5073, 3032, 3426, 2793, 2158, 4464, 3455}

title           Touched With Fire
subtitle        not found
authors         Kay Redfield Jamison
categories      Psychology
average_rating  4.01
num_pages       384.0
id              4100
Name: 4100, dtype: object

```

OUTPUT FOR INDEXING BY AUTHOR:

```

marilynne 0
robinson 0 26 198 402 416 1502 1572 1588 3352 3486 3512 3547 3625 5180 5854 6595
charles 1 162 194 267 371 428 445 780 787 813 948 956 1047 1180 1194 1200 1238 1329 1373 140
2 1577 1606 1613 1861 2065 2105 2141 2147 2253 2302 2304 2305 2366 2417 2437 2507 3063 3067
3136 3194 3216 3217 3259 3273 3317 3538 3539 3540 3549 3701 3778 3910 3920 4201 4203 4216 42
28 4268 4339 4676 4755 4860 4978 5079 5105 5122 5180 5392 5423 5437 5469 5470 5579 5689 5692
5721 5816 6012 6055 6226 6333 6364 6452 6569 6573 6649 6718
osborne 1 1223 2166 2172 2173 4077 4867 4868 4869 4871
agatha 1 6 17 25 40 41 44 45 46 47 48 49 50 51 52 53 54 55 56 57 72 74 75 88 1492 1611 1620
2517 2552 2619 6204 6205 6206 6207 6208 6209 6210 6211 6212
christie 1 6 17 25 40 41 44 45 46 47 48 49 50 51 52 53 54 55 56 57 72 74 75 88 1492 1611 162
0 2231 2517 2552 2619 3600 4432 6204 6205 6206 6207 6208 6209 6210 6211 6212
stephen 2 8 634 637 786 1155 1614 1669 1756 1758 1823 1842 1851 1861 2045 2186 2242 2261 236
5 2381 2386 2387 2390 2391 2392 2415 2436 2544 2582 2795 2862 2873 2965 2966 2967 2969 2970
2973 2974 2977 2978 2979 2980 2981 2982 2985 2988 3006 3077 3108 3144 3272 3366 3367 3382 34
00 3401 3404 3409 3417 3419 3456 3469 3498 3525 3562 3636 3638 3686 3687 3702 3725 3789 3814
3826 3835 3836 3840 3844 3859 3865 3886 3899 3922 4217 4221 4243 4252 4349 4364 4382 4387 44
10 4420 4428 4433 4447 4550 4552 4558 4586 4660 4661 4697 4890 4922 5151 5311 5327 5525 5542
5557 5629 5736 5790 5897 6100 6118 6119 6149 6157 6227 6252 6314 6427 6653

```

OUTPUT FOR SEARCHING BY AUTHOR:

Retrieves all the Books related to the query
Stanley(Author of a Book)

```
////////////////////////////////////
stanley
////////////////////////////////////
stanley 26 167 519 635 1413 1572 1588 1835 2167 2168 3192 3198 3512 3547 3820 4256 5180 6230
6475

{1413, 519, 26, 4256, 1572, 167, 1835, 1588, 3512, 5180, 3192, 6475, 6230, 3547, 3820, 2167,
2168, 635, 3198}

title           The Spy Who Barked
subtitle        not found
authors         George Edward Stanley
categories      Juvenile Fiction
average_rating  3.47
num_pages       48.0
id              1413
Name: 1413, dtype: object

title           What Would Machiavelli Do?
```

OUTPUT FOR RELEVANCE FEEDBACK:

Retrieves all the ids and the author name of the books so that user can choose the relevant one's by giving their id's.

1413	George Edward Stanley
519	Stanley Bing
26	Kim Stanley Robinson
4256	Edward C. Klatt;Stanley Leonard Robbins;Vinay Kumar
1572	Kim Stanley Robinson
167	Stanley Bing
1835	Erle Stanley Gardner
1588	Kim Stanley Robinson
3512	Kim Stanley Robinson
5180	Kim Stanley Robinson;Arthur Charles Clarke
3192	Pedro Calderón de la Barca;Stanley Appelbaum
6475	Conner Gorry;David Stanley
6230	Stanley Wiater;Christopher Golden;Hank Wagner
3547	Kim Stanley Robinson
3820	Thomas J. Stanley;William D. Danko
2167	George Edward Stanley
2168	George Edward Stanley
635	Darlene Clark Hine;William C. Hine;Stanley Harrold
3198	Pedro Calderon de la Barca;Stanley Appelbaum

OUTPUT FOR RELEVANCE FEEDBACK:

User gave id's 26 and 1572 as relevant one's and it is stored as a dictionary.

```
Out[24]: {'stanley': ['26', '1572']}
```

OUTPUT FOR RELEVANCE FEEDBACK:

Whenever another user comes up with the same query feedback of previous user is considered and books with id 26 and 1572 are kept as top 2 and rest books followed by that ensuring Relevance Feedback.

```
title           The Years of Rice and Salt
subtitle        not found
authors          Kim Stanley Robinson
categories       Black Death
average_rating   3.73
num_pages        772.0
id               26
Name: 26, dtype: object
title           The Memory of Whiteness
subtitle        A Scientific Romance
authors          Kim Stanley Robinson
categories       Fiction
average_rating   3.56
num_pages        352.0
id               1572
Name: 1572, dtype: object
title           The Spy Who Barked
subtitle        not found
authors          George Edward Stanley
```

THANKYOU!

