<u>הקדמה:</u>

סט הנתונים שאיתו בחרנו לעבוד הוא קורפוס של ספרים מ4 ז'אנרים עיקריים:

1. הרפתקאות
2. אוטוביוגרפיות
3. מדעי המדינה ופילוסופיה
4. נובלים

אנחנו נותנים לכל ספר מזהה ייחודי ושומרים את הקישור הזה בdictionary, דוגמה עם 21 ספרים:

```
{1: 'Data/Bushido the Soul of Japan.txt',
 2: 'Data/The Wonderful Wizard of Oz.txt',
 3: "Data/Alice's Adventures in Wonderland.txt",
 4: 'Data/Theologico Political Treatise.txt',
 5: 'Data/The Wit and Humor of America Volume IX.txt',
 6: 'Data/Political Ideals.txt',
 7: 'Data/The Autobiography of Benjamin Franklin.txt',
 8: 'Data/The Adventures of Tom Sawyer.txt',
 9: "Data/Pride and Prejudice a play founded on Jane Austen's novel.txt",
 10: 'Data/Autobiography of Makataimeshekiakiak or Black Hawk.txt',
 11: 'Data/The Call of the Wild.txt',
 12: 'Data/The Man in the Brown Suit.txt',
 13: 'Data/The Prince.txt',
 14: 'Data/The Secret Garden.txt',
 15: 'Data/Treasure Island.txt',
 16: 'Data/Readings on Fascism and National Socialism.txt',
 17: 'Data/Heart of Darkness.txt',
 18: 'Data/The Writings of Thomas Paine Volume 4 1794 to 1796 The Age of Reason.txt',
 19: 'Data/Autobiography of Benjamin Franklin.txt',
 20: 'Data/Up from Slavery An Autobiography.txt',
 21: 'Data/Beautiful Joe An Autobiography.txt'}
```

<u>ניקוי הספרים:</u>

אנחנו קוראים כל ספר כקובץ טקסט והופכים אותו לוקטור ארוך של מילים.
עבור כל ספר אנחנו הופכים את כל האותיות לאותיות קטנות, מורידים תווים שלא מהאלפא בית האנגלי ומסננים stop words.

<u>בניית הטבלאות הראשונית:</u>

לכל ספר אנחנו בונים את הטבלות:
Inverted Index – טבלה ששומרת לכל מילה רשימה של הספרים בהם היא מופיעה:

```
+---------------+--------------------+
|          words|               docID|
+---------------+--------------------+
|            296|                [19]|
|      ammonites|                 [4]|
|antisocialistic|                [16]|
|  apprehensions|     [15, 19, 7, 11]|
|      arguments|[18, 19, 21, 1, 3...|
|            art|[13, 14, 15, 16, ...|
|        attackd|             [19, 7]|
|        barrier|             [12, 9]|
|          besom|             [14, 1]|
|         biting|[15, 17, 18, 21, ...|
|         blairs|                [12]|
|         bleeve|                 [8]|
|        blossom|             [20, 1]|
|       bowsprit|                [15]|
|       brackets|                 [7]|
+---------------+--------------------+
only showing top 15 rows
```

tf table – טבלה ששומרת עבור כל מילה את ערך הtf שלה בכל ספר לפי הנוסחה שלמדנו בכיתה:

- The log frequency weight of term t in d is defined as follows

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

הטבלה עצמה:

```
+---------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
|          words|book_1_tf|book_2_tf|book_3_tf|book_4_tf|book_5_tf|book_6_tf|book_7_tf|book_8_tf|book_9_tf|
+---------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
|            296|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|
|      ammonites|      0.0|      0.0|      0.0|  1.60206|      0.0|      0.0|      0.0|      0.0|      0.0|
|antisocialistic|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|
|  apprehensions|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0| 2.146128|      0.0|      0.0|
|      arguments|      1.0|      0.0|1.4771212|2.0791812|  1.69897|2.5563025|2.6232493|      0.0|      0.0|
|            art|2.1760912|  1.60206|      0.0|  1.60206|2.4771214|2.4771214|2.7993405|  2.60206|      0.0|
|        attackd|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0| 2.447158|      0.0|      0.0|
|        barrier|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|1.9542425|
|          besom|      1.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|
|         biting|      1.0|  1.30103|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|      0.0|
+---------------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
only showing top 10 rows
```

counter table – טבלה ששומרת עבור כל מילה את המונה שלה בכל ספר:

words	book_1	book_2	book_3	book_4	book_5	book_6	book_7	book_8	book_9	book_10	book_11	book_12	book_13
07	0	0	0	6	0	0	0	0	0	0	0	0	0
296	0	0	0	0	0	0	0	0	0	0	0	0	0
abruptness	0	0	0	0	0	0	0	0	8	0	0	0	0
accumulation	0	0	0	0	0	0	0	0	8	9	10	11	0
ammonites	0	0	0	0	4	0	0	0	0	0	0	0	0
antisocialistic	0	0	0	0	0	0	0	0	0	0	0	0	0
apprehensions	0	0	0	0	0	0	0	14	16	0	0	0	144
archduchy	0	0	0	3	0	0	0	0	0	0	0	0	0
arguments	0	0	2	66	12	5	36	42	16	9	0	0	96
art	0	2	0	81	4	30	30	63	16	45	80	77	372

only showing top 10 rows

idf + counter table – טבלה ששומרת עבור כל מילה את ערך הidf שלה ביחס לכל הקורפוס ואת כמות הספרים שהיא מופיעה בהם סך הכל:

counter	idf	words
1	1.3222192947339193	296
1	1.3222192947339193	ammonites
1	1.3222192947339193	antisocialistic
4	0.7201593034059569	apprehensions
10	0.3222192947339193	arguments
16	0.11809931207799448	art
2	1.021189299069938	attackd
2	1.021189299069938	barrier
2	1.021189299069938	besom
7	0.47712125471966244	biting
1	1.3222192947339193	blairs
1	1.3222192947339193	bleeve
2	1.021189299069938	blossom
1	1.3222192947339193	bowsprit
1	1.3222192947339193	brackets
1	1.3222192947339193	brands
1	1.3222192947339193	buggies
1	1.3222192947339193	captainover
1	1.3222192947339193	carnegie
9	0.36797678529459443	cautious

only showing top 20 rows

tf-idf table – הטבלה הסופית בה הכפלנו את ערך הtf של כל ספר בערך הidf שחישבנו בסעיף הקודם לפי הנוסחה שנלמדה בכיתה:

- The tf-idf weight of a term is the product of its tf weight and its idf weight.

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

הטבלה עצמה:

words	tf_idf_book_1	tf_idf_book_2	tf_idf_book_3	tf_idf_book_4	tf_idf_book_5
296	0.0	0.0	0.0	0.0	0.0
ammonites	0.0	0.0	0.0	2.1182745909156515	0.0
antisocialistic	0.0	0.0	0.0	0.0	0.0
apprehensions	0.0	0.0	0.0	0.0	0.0
arguments	0.3222192947339193	0.0	0.47595696223668355	0.6699522980531204	0.5474409023538536
art	0.256994873048431	0.1892021792268405	0.0	0.1892021792268405	0.29254632774065664
attackd	0.0	0.0	0.0	0.0	0.0
barrier	0.0	0.0	0.0	0.0	0.0
besom	1.021189299069938	0.0	0.0	0.0	0.0
biting	0.47712125471966244	0.6207490850112853	0.0	0.0	0.0

only showing top 10 rows

מימוש האלגורתמים:

Cosine similarity between every pairs of books –

Top five:

1) Autobiography of Benjamin Franklin.txt and The Autobiography of Benjamin Franklin.txt is: **0.9775778571039143**

   Short explanation: 2 distributions of the same book that edited by different peoples

2) The Wit and Humor of America Volume IX.txt and The Man in the Brown Suit.txt is: **0.7777794172660059**

   Short explanation: Old books (around 1920) wrote with the same linguistic style.

3) Heart of Darkness.txt and Treasure Island.txt is: **0.7685012032602614**

   Short explanation: Adventures stories about voyage around the world.

4) Beautiful Joe An Autobiography.txt and The Wit and Humor of America Volume IX.txt is: **0.7669732708700917**

5) The Man in the Brown Suit.txt and Heart of Darkness.txt is: **0.7530038563738879**

Short explanation: Mystery and crime plots with sophisticated characters

Cosine similarity between query and books-

Few examples for interesting query made

Query: "why the kingdom of darius, occupied by alexander, did not rebel against the successors of the latter after his death?"

Short description: this query is subject inside one of our books in the corpus (the prince)

Best results:
The Prince.txt the score is: **0.6911683954044678**
The Writings of Thomas Paine Volume 4 1794 to 1796 The Age of Reason.txt the score is: **0.27372709840115766**
Bushido the Soul of Japan.txt the score is: **0.20862201172069114**

Query: "Bushido the Soul of Japan"

Short description: title of the book

Best results:
Bushido the Soul of Japan.txt the score is 1.2696520570566676

Query: "black people history"

Short description: general query

Best results:
Up from Slavery An Autobiography.txt the score is: **1.222529982481461**
Readings on Fascism and National Socialism.txt the score is: **1.5168521476006138**
Autobiography of Makataimeshekiakiak or Black Hawk.txt the score is: **1.2412319176444084**

Query: "what is Fascism and National Socialism?"

Short description: the main subject of the book

Best results:
Readings on Fascism and National Socialism.txt the score is: **2.027376062639929**
Political Ideals.txt the score is: **0.28783055933019197**

Query: "the mad hatter"

Short description: character in the book

Best results:
Alice's Adventures in Wonderland.txt the score is: 0.2674732553300295

Kmeans –
As we mentioned at the beginning, we separated the books to 4 main subjects:
Adventure, Autobiography, Novels and Political philosophy.
This is how our final separation algorithm went:

Cluster 1:
['The Writings of Thomas Paine Volume 4 1794 to 1796 The Age of Reason.txt', 'Autobiography of Benjamin Franklin.txt', 'Up from Slavery An Autobiography.txt'],

Cluster 2:
['Autobiography of Makataimeshekiakiak or Black Hawk.txt'],

Cluster 3:
['Bushido the Soul of Japan.txt', 'The Wonderful Wizard of Oz.txt', "Alice's Adventures in Wonderland.txt", 'Theologico Political Treatise.txt', 'The Wit and Humor of America Volume IX.txt', 'Political Ideals.txt', 'The Autobiography of Benjamin Franklin.txt', 'The Adventures of Tom Sawyer.txt', "Pride and Prejudice a play founded on Jane Austen's novel.txt", 'The Call of the Wild.txt', 'The Man in the Brown Suit.txt', 'The Prince.txt', 'Readings on Fascism and National Socialism.txt', 'Heart of Darkness.txt'],

Cluster 4:
['The Secret Garden.txt', 'Treasure Island.txt', 'Beautiful Joe An Autobiography.txt']

We can see that most of the adventure books were classified together were as the Autobiography and Political philosophy separation were not a full success.