שם הקורס: מבוא לאנליזה של ביג דאטה

מגישים: מתן רחמני (311449615), דניאל וייסרברג (311130314), יעל ליבנה (305545527(

**הקדמה:**

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:

A screenshot of a cell phone

Description automatically generated

**ניקוי הספרים:**

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.

**בניית הטבלאות הראשונית:**

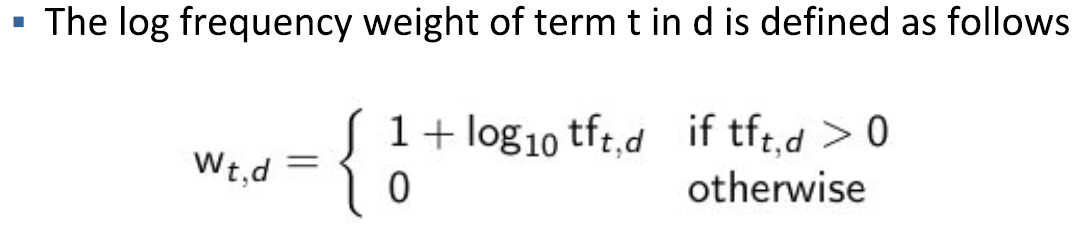
לכל ספר אנחנו בונים את הטבלות:

Inverted Index – טבלה ששומרת לכל מילה רשימה של הספרים בהם היא מופיעה:

A close up of text on a white background

Description automatically generated

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



הטבלה עצמה:

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

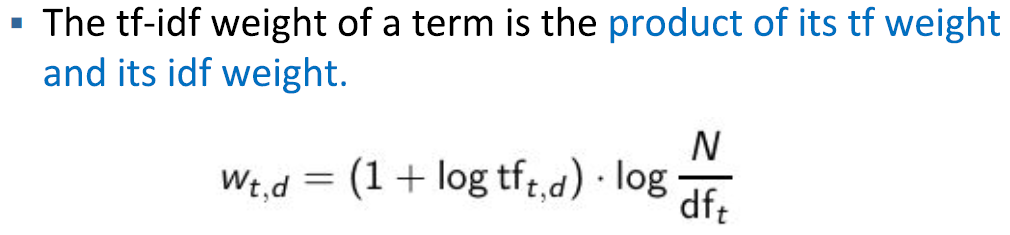
Description automatically generated

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

A close up of text on a white background

Description automatically generated

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



הטבלה עצמה:

A screenshot of a cell phone

Description automatically generated

**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

1. 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.

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

**Short explanation:** Adventures stories about voyage around the world.

1. Beautiful Joe An Autobiography.txt and The Wit and Humor of America Volume IX.txt is: **0.7669732708700917**
2. 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.tx**t 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.**