

*אחזור מידע 2021-2020*

*צוות הקורס : ד״ר ניר גרינברג*

*סהר בריבי, עמית ליבנה ובוריס סובול*

**פרויקט תכנות**

**דו"ח בניית מנוע לאחזור מסמכים – חלק א'**

1. **עיצוב התוכנה:**
   1. מטרת חלק א' הינה בניית אינדקס הופכי עבור מאגר מסמכים, באמצעות עיבוד מקדים למאגר. ראשית המנוע "**קורא"** (reader) את כל מאגר טוויטים שקיבלנו ( כ 10 מליון טוויטים = 2Gb). שנית ביצענו עבור כל מסמך שקראנו את שלב ה"**פרסור**" (parser), שבו הפרדו את הטקסט לאוסף של מונחים (terms) וכל term עובר את הדרישות שנקבעו. בשלב זה הוסרו כל מונחים מסוג words-stop . כעת, נעבור לשלב ה"**אינדוקס**" (indexer) שבו אנחנו יוצרים את המילון ואת ה Posting.וכעת נעשה את שלב ה"**חיפוש**" (search) שבו בעזרת פונקציות דמיון glove נבצע הרחבת שאילתה ונחפש מילים דומות בעזרת מודל מאומן. לבסוף, בשלב "**הדירוג**" (ranker) מדרגים את כל ה- s'ID שקיבלנו ע"י שימוש בוקטורים שיצרנו לכל מסמך בעזרת מודל ה-GloVe, ולאחר מכן שימוש בחישוב cosine similarity בין כל אחד מוקטורי המסמכים לבין וקטור השאילתא.

**מחלקת reader –** מטרתה לקרוא את כל הData בפורמט parquetולהחזיר אובייקט שמכיל רשימה של כל המסמכים העומדים להחלץ ולעבור פירסור.

**מחלקת parser\_module –** מטרתה לפרק כל מסמך(טוויט) לאוסף של terms לפי סט חוקים. נבצע פרסור רק לfull text וURL שקיים בכל מסמך. המחלקה רצה בלולאה על מסמך מסמך (לפי ואחרי Stemmer ) ומבצעת חוקים ומורידה מהטקסט את stop-word. נציין שהחלפנו את כתובת URL הקצרה שנמצאת במסמך עם כתובת URL חיצוני אם קיימת, כמו כן אם לא היה קיים URL קצר אבל חיצוני הוספנו אותו לטקסט. עבור כל מסמך הוספנו לו שדות 2 שבו קיים מילון ( אחד עבור מילים רגילות והשני עבור מילים שעברו Stemming) של terms ומספר הופעות המילה במסמך. בנוסף, שמרנו 2 מילונים (עם ובלי Stemming) של כל המילים במאגר וכמות הפעמים שהופיעו. בנוסף על כך, שמרנו מילון עבור כל הישיות.

**תוספות חשובות במחלקה:**

* הוספת מילים לstop-words על מנת לצמצם את המונחים במילון לרלוונטים:

{"www", "https", "http", "status", "instagram.com", "twitter.com", "t.co", "rt", "RT",'instagram','facebook','twitter','co','com','web'}

מילים אלה מופיעות מספר רב של פעמים ואינם מספקות מידע נוסף. לכן, על מנת לצמצם את נפח המילון בחרנו להוסיף אותם לstop-words.

* החלפת ה-URL המקוצר שנמצא בתוך הציוץ בURL הרחב והמפורט יותר: במידה ולא קיים קצר, הוכנס המורחב אל סוף הציוץ. במידה וקיימים יותר מ-1 בשדה המורחב, הוחלף הקצר בראשון והשאר הוכנסו בסוף.
* הורדנו את כל הטוויטים שהם ציוצים חוזרים. לצורך העניין, כל הציוצים המתחילים בRT. הורדה זו התבצעה כיוון שציוצים אלו לא מוסיפים מידע חדש לקורפוס ובנוסף רק מכבידים על גודל הinverted index ועל כל תהליך יצירתו.

**מחלקת stemmer –** מטרתה לבצע stemming על כל terms. השתמשנו בקוד פתוח, בחבילה .LancasterStemmer

**מחלקת document -** מייצגת את כל השדות של המסמך.

**מחלקת indexer –** מטרתה לבצע יצירת האינדקס ההופכי, כאמור יצירת המילון וקבצי Posting. המחלקה מקבלת את המילים מהparser ומתחילה לבנות את האינדקס ההופכי. בנוסף, בשלב זה, אנחנו עוברים על המילון מילים ששמרנו בשלב הparser ומבצעים עליו את החוק של אותיות גדולות קטנות ביחס לכל המילים במילון. נעבור גם על מילון ישיות ונוריד ישיות שמופיעות פעם אחת בלבד. במילון Posting file עבור כל מילה – נשמור : מזהה יחודי id twitter, כמות הופעות של המילה במסמך Tf, ו כמות הופעות במסמך חלקי כמות הופעות מקסימלי במסמך tf / max(tf) .

נוסף לכך, שמרנו מילון שעבור כל מילה – נשמור : מספר הופעות בכל המילון, כמות מסמכים שהופיע Df, אות ראשונה של המילה, שהיא בעצם הפוינטר לקובץ פוסטינג. חוץ מזה שמרנו גם מילון עבור כל מסמך Id – נשמור: כמות הופעות מקסימלי במסמך max(tf), מספר מילים יחודיות, ואורך המסמך.

באופן זהה נשמור גם למילים שעברו Stemmer כל אחד משלושת המילונים.

**תוספות חשובות במחלקה:**

* לאחר שהמסמך עובר אינדוקס(מחוץ למחלקה), אנו שומרים את המילונים שבו(עם ובלי סטמינג) באובייקט מסוג מילון, אשר מפתחותיו מסודרים עפ"י ארבעת הספרות האחרונות בid של אותו המסמך שמוכנס. זאת על מנת ליצור קבצים קטנים מבחינת הנפח לטובת ביצוע דירוג המסמכים שמצריך את המילונים של כל מסמך רלוונטי.
* הסרת מונחים שפעם אחת – ההסרה מתבצעת כיוון שמונחים אלו הינם חריגים בקורפוס, משמע לא יוסיפו מידע חשוב וכן הם כנראה אינם מופיעים במודל המאומן(של GloVe).

**מחלקת searcher -** מטרתה למצוא מילים דומות לשאילתה בעזרת פונקציות דמיון GloVe . המחלקה מקבלת את השאילתה מהמשתמש ולאחר מכן אנו מרחיבים את השאילתא, כלומר מוסיפים לחיפוש עוד מונחים אשר יכולים להוסיף עוד מידע ולחזק את הדירוג של הטוויטים, בנוסף בשלב הזו אנו מייצרים וקטור אשר מייצג את כל המונחים של השאילתא לאחר פירסור, על מנת שנוכל להשתמש בו בשלב הדירוג. בעזרת קובץ מאומן שקיבלנו מקוד פתוח מחבילת genism glove2word2vec, אנו מרחיבים את השאילתא בעזרת פונקצית דמיון ווקטורית.

לאחר שיצרנו רשימה של כל המילים שאנחנו מעוניינים לחפש, אנו משתמשים בקבצי הפוסטינג על מנת לאתר את כל הids הרלוונטים לכל מילה.הוספת הids היא לתוך מילון אשר יכיל את כל הids ואת כמות הפעמים שהופיעו בזמן החיפוש של כל מונח.

לבסוף, נחזיר את ה-2000 Ids אשר הופיעו הכי הרבה פעמים בסדר יורד.

**מחלקה ranker –** מטרתה לדרג את כל המסמכים שקיבלנו מהשלב הקודם. המחלקה מקבלת את הids של הציוצים הרלוונטים ושולפת מהדיסק את המונחים של כל ציוץ ומכניסה לרשימה. לאחר מכן כל רשימה של מונחים הופכת לוקטור כך שלבסוף אנו מקבלים רשימה של וקטורים שהם מייצגים את המונחים של כל ציוץ. לביצוע הדירוג אנו משתמשים בחישוב של cosine similarity אשר מקבל שני וקטורים ומחזיר את רמת הדימיון ביניהם בין 0 ל-1. מבצעים את החישוב לכל וקטור ברשימה אל מול הוקטור של השאילתא שיצרנו בשלב החיפוש. את התוצאות של החישובים אנו ממיינים מהדימיון הגבוה ביותר לנמוך, כל שנוכל לקבל את הציוצים הדומים ביותר לשאילתא בתחילת הרשימה של התוצאות.

**מחלקת search\_engine –** מטרה לנהל את כל מהלך איחזור המידע. מרגע קריאת המידע, פירסורו ואינדוקסו, כמו כן, שמירה לדיסק ועד לרגע קבלת השאילתא והחזרת המסמכים הרלוונטים מדורגים על פי קירובם.

* 1. **דרכי התמודדתם עם מגבלת הזיכרון** של המחשב והפעולות שנקטתם על מנת להביא לזמן ריצה מיטבי: על מנת לא להעמיס על הזיכרון הראשי ובכדי לנצל אותו בצורה הכי יעילה שאפשר, חילקנו את המידע בכל חלקי התהליך. ראשית, בקבלת קבצי המידע, קראנו אותם בנפרד ועל כל חלק ביצענו את המשך ההליך. שנית, בזמן הפירסור עיבדנו כמות מידע שאיתו הזיכרון הראשי יכול היה להתמודד ובסוף העיבוד ייצאנו את המידע אל תוך קבצים בכונן הקשיח, על מנת לפנות זיכרון לחלק המידע הבא. שלישית, עם תחילתו של הליך האינדוקס, את הקובץ הגדול שמכיל את המונחים של כל מסמך, ייבאנו בחלקים ועבדנו עם כל חלק בנפרד על מנת לא להעמיס על הזיכרון הראשי. בסוף תהליך האינדוקס שמרנו את המידע(הקובץ ההופכי) שוב פעם בחלקים כפי שעשינו בתהליך הפירסור. חשוב לציין שלאחר העברת המידע לאובייקט מסויים, נמחק המידע מהאובייקט המעביר לטובת פינוי הזיכרון.
  2. **אופן שמרת את קבצי ה-Posting ,סוג הקבצים, כמות הקבצים, מה מכיל כל קובץ וכדומה.** – שמרנו את הקבצי Posting בפורמט pkl, עם ובלי Stemmer, כאשר הקבצים מחולקים לאותיות לפי אות הראשונה של כל מילה (קיים חשיבות לאות גדולה וקטנה), כמו כן גם שמרנו קבצים עבור מונחים אחרים כמו מספרים, סימנים וכו'.

כמות הקבצים – **320** קבצים. כל קובץ מכיל מילונים בגדלים שונים, אשר מפתחותיהם הינם המונחים. לכל מונח קיים מילון אשר מפתחותיו הינם הids של הציוצים הרלוונטים לכל מונח. לכל id שמרנו את tf ואת החישוב של tf/tf\_max.

\* צורת שמירה זו איפשרה לנו להקל על צורת חיפוש הציוצים הרלוונטים.

\*חשוב לציין שבכדי לאתר את כל הids הרלווטים לאותו מונח צריך לבצע חיפוש של המונח בכל החלקים של אותו קובץ.

* 1. **עליכם לציין את הסיבות לבחירת גודל קבוצת הטוויטים החלקית וכן להציג תיעוד תהליך יצירת הקבצים ההופכיים (מילון וקובץ posting).** בחרנו לחלק את קבוצת הטוויטים למקסימום 500 אלף טוויטים בכל קבוצה(כיוון שיש קבצים שמכילים גם פחות וכן ציוצים שנמחקים) כיוון שבחלוקה זו הזיכרון הראשי יכול היה להתמודד מבלי לקרוס וכך לבצע את שאר ההליך בצורה מהירה ויעילה.

תיעוד התהליך-

* + 1. נעבור על כל חלק ששמרנו בתוך הקובץ לאחר הפירסור. אשר יכול להכיל עד 500 אלף מסמכים.
    2. נאנדקס כל מסמך שמתקבל וניצור עבור כל מונחיו את קבצי הפוסטינג הרלוונטים ונערוך את המילון בהתאם(במידה וצריך נמחוק מהמילון).
    3. כפי שעשינו גם בשלב הפירסור, לאחר מקסימום 500 אלף מסמכים נשמור את קבצי הפוסטינג אשר מחולקים עפ"י אותיות ראשונות.
    4. לאחר שסיימנו לעבור על כל המסמכים, נוכל לקבל מילון מוגמר אשר עבר את כל השינויים שהיה צריך, לכן נשמור אותו רק בסוף תהליך האינדוקס בדיסק.
  1. **עליכם לציין את פריטי האינפורמציה הנוספים ששמרתם ולהסביר מדוע שמרתם דווקא אותם.**
* **חוק הקורנה-וירוס –** שמנו לב שיש הרבה מילים נרדפות בקופרוס למילה קורנה וירוס. לכן, בחרנו לשמור את האינפורמציה תחת מונח אחד coronavirus. כך נוכל להתחשב בו בצורה אחידה מבחינת הקורפוס ושנקבל את כמות הופעות של הביטוי בקורפוס (Df).

אופן מימוש: עבור כל term נבדוק אם הוא אחד מכל הביטויים שלעיל Covid-19'' COVID-19' 'Covid' 'COVID' 'covid-19' 'covid19' 'Covid19' 'COVID19' במידה וכן נחליף אותו בטקסט ב coronavirus.

* **חוק מסתיים ב`s או s'** – שמנו לב שמילים מסוג term+`s מופיעות בהרבה מסמכים ולא מטופלות. בחוק זה נשמיט מהterm את הtokens s . בצורה זו נייעל את תהליך אחזור ובנוסף נקטין את נפח המילון. מנגד מילים שמכילות term+‘ אבל לא s (לדוגמא don`t) לא נסיר מהמילון.

אופן מימוש: עבור כל term נבדוק אם המילה מסתיימת ב`s או s' , במידה וכן נשמור את המילה החדש ללא `s או s'.

* 1. **שני החוקים שהוספתם, יש להדגים כיצד החוקים באים לידי ביטוי בשני טוויטים שונים ב-Dataset:**

**חוק הקורנה-וירוס:**

* בID מספר - 1288840098893660160':

“@ScottAdamsSays: anyone ever seen story guy got COVID-19,

got early treatment combo HCQ, zinc, Azithro…”

החלפנו COVID-19 ב coronavirus.

* בID מספר - '1288839269318942726'

” '\'So far, so good\' on leading COVID vaccine, says AstraZeneca https://t.co/ztKkYVdK6J https://t.co/gvgQKwfXRX'”

החלפנו COVID ב coronavirus.

**חוק מסתיים ב`s:**

* בID מספר - '1288840340992962560':

" 'RT @Tracie0313: Yes!!! MY doctor should be allowed to prescribe medication based on my symptoms and if there’s even a possibility this drug…'

שונה המונח there’s ל there.

* בID מספר - ' '1288832516074266629:

" 'RT : Nothing will stop me from casting my vote for Joe Biden -- COVID-19, long lines, mail delays, trump\'s Gestapo goons...NOTHI…'

שונה המונח trump\'s ל trump.

* 1. **עבור כללים/ חוקים נוספים שהוספתם (ב-Parser וב-Indexer (יש להסביר את הרעיון של כל חוק וכיצד מימשתם אותו** :
* **אורך המסמך -** בחרנו לשמור את האינפורמציה הזו, כי במודל BM25 משפטים ארוכים יותר מקבלים דירוג נמוך לעומת משפטים קצרים המקבלים דירוג גבוה יותר. שמרנו את אורך המסמך לפני שביצענו עליו את החוקים והסרה של מונחים שאינם עומדים בדרישות.

אופן מימוש: לאחר שלב אינדוקס של כל מסמך, הוספנו במילון המיועד למידע אודות מסמך: document info את אורך המסמך.

* **tf/tfmax -** אינפורמציה זו מסייעת לנו באלגוריתם דירוג מסמכים.

אופן מימוש : בשלב אינדוקס כשאנחנו עוברים מסמך אנחנו שומרים את כמות הפעמים שהופיעה הכי הרבה פעמים(tfmax). עבור כל term מהמסמך אנחנו שולפים את כמות הפעמים שהופיע המילה במסמך(tf) . כעת נשמור את tf/tfmax עבור כל term בקובץ posting.

הערה:אמנם חוקים אלה שהוספנו למסמך אינם מועילים לנו בשלב זה, אך בתור ראייה עתידית לשלב הבא של הפריוקט חשבנו לשמור מידע זה על מנת להקל עלינו בשליפה של נתונים אלה.

* 1. **הסבירו בצורה מפורטת את האלגוריתם בו השתמשתם במחלקת Ranker:**

לאחר קבלת ids הציוצים הרלוונטים, ייבאנו מתוך הכונן את המילון הרלוונטי(עם או בלי סטמינג) והוספנו אותו לרשימה שיכיל את כל המילונים. משם יצרנו לכל מילון של ציוץ וקטור בעזרת המודל המאומן של GloVe.

על מנת לחשב את הקירובים השתמשנו בפונקציית החישוב של cosine similarity, אשר מקבלת כקלט 2 וקטורים ומחשבת את הקירוב ביניהם. הפונקציה מכפילה בין שני הוקטורים ומחלקת במכפלת הוקטורים המנורמלים. בצורה כזו שלחנו את כל הקטורים מהרשימה אל הפונקציה ולבסוף קיבלנו רשימה ממויינת של הקירובים והציוצים.

* 1. **השתמשנו בקוד פתוח הבא:** -
* מחבילת nltk.stem השתמשנו SnowballStemmer

קישור למקור: <https://www.nltk.org/api/nltk.stem.html>

* מחבילת genism השתמשנו glove2word2vec , KeyedVectors

קישור למקור : https://nlp.stanford.edu/projects/glove/

* 1. **מידע נוסף להבנת התכנית:**
* בפרסור URL לאחר שאנחנו מפרסרים את הכתובת לפי החוקים שהוגדרו במסמך דרישות אנחנו מסירים את כל המילים אשר מכילות אותיות ומספרים יחד. בחרנו להוריד אותם מכיוון שהם לא עוזרים לנו בתהליך האחזור מכיוון שאינם מספקים מידע אודות הציוץ לכן רק יכולים לגרוע מבחינת זיכרון ואחזור.
* במהלך עבודה על הפריוקט יצרנו חוקים על מנת לבצע את האחזור בצורה יעילה ונכונה. חוק אימוגי חוק הציטוטים אשר מיפנו אותם ושמרנו אותם. כאשר התחלנו את של הבדיקות ראינו שחוקים אלה מאד מכבידים לנו על המילון, ועל זמני הריצה ולכן, לפי ראות דעתנו החלפנו אותם בחוקים אחרים.כמו כן, בשלב הפירסור שמרנו את האינדקס של כל מילה במיקום שהיא מתחילה במסמך, אבל גם במקרה הזה בגלל שיקולי מקום וזמן החלפנו חוק זה.
* אנחנו מסירים את כל האשטגים אשר מכילים סימנים או שפה אחרת מאנגלית. כמו למשל #קורנה. מונח זה לא ישמר, מכיוון שאינו עוזר לנו בתהליך אחזור.
* אם ישות מופיעה במילון הישויות פעם אחת, הישות תימחק מהמילון הגדול וגם מהמילון של אותו מסמך. בנוסף, כאשר יש התנגשות בין חוק האותיות הגדולות לחוק הישויות אנחנו משאירים את הישויות בצורה המקורית שלהן.
* בחרנו להוריד מהמילון את כל המונחים אשר הופיעו פעם אחת, מכיוון שמונחים אלה אינם שכיחים ולא מועילים לנו בשלב האחזור ורק מכבידים בזיכרון נוסף. יתרה מכך לרוב למילים אלה לא נמצא ווקטור במודל המאומן של הGlove.
  1. **שלושה יתרונות ושלושה חסרונות של המנוע של המנוע שלכם על פני מנועים אחרים:**

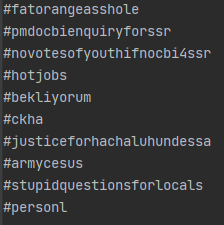
**יתרונות:**

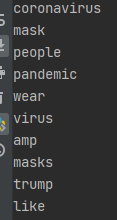
* מכיוון שאנחנו מפצלים את הקבצי Posting לאחר מקסימום 500,000 מסמכים שעברו אינדוקס, אנחנו בעצם מקלים על הכתיבה לזיכרון, ומורידים העמסה מהזיכרון הראשי.
* בגלל שאנחנו מחלקים את קבצים שמכילים את המידע הדרוש לאחזור המסמכים ולביצוע הדירוג (המילונים האישים עבור כל מסמך) ל1000 קבצים שבו כל מסמך ממופה על פי 4 הספרות האחרונות שלו. כתוצאה מכך נוכל לגשת לקובץ מסיום לפי סיומת ID במקום לעבור על כולם. כמובן, שכתוצאה מכך נקל על תהליך אחזור מבחינת מעבר מסמכים שהמילון מכיל ונייעל את המהירות.
* אנחנו מסווגים את הקבצי Posting לפי אות הראשונה או לפי הסימן. אופציה זו מקלה עלינו לשלוף מידע מהקבצים בצורה מהירה ויעילה יותר.

**חסרונות:**

* מכיוון שאנחנו מפצלים את הקבצי Posting לאחר 250,000 קבצים, אז קיימים לנו המון קבצים עבור כל ערך. כ40 Posting עבור אות בודדת, מכאן אנחנו צריכים לעבור על כולם על מנת למצוא את הערך.
* בחיפוש אני מרחיבים את השאילתה ע"י דמיון מסמכים המאומן של GloVe מה שעלול להחזיר לנו מונחים גם שלא בהכרח רלוונטים.
* שמירת ישיות – מכיוון שאנחנו לא משתמשים בספריית ניתוח ישיות של nltk, אלא בRegex שיצרנו אז אנחנו עלולים לסווג ישיות למרות שבפועל אינן ישיות. לדוגמא Thursday July יסווג לישות למרות שבפועל אינו ישות.

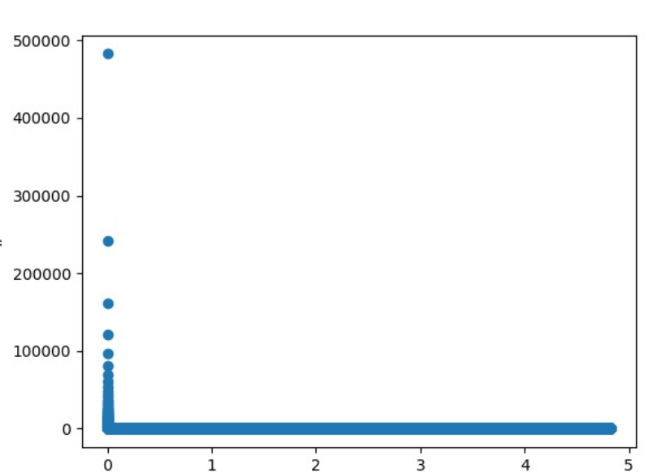
1. 1. מונחים שונים לפני סטמינג: **4826546**
   2. מונחים שונים אחרי סטמינג: **4560936**
   3. כמה terms שונים שהם מספרים יש במאגר: **1098471**

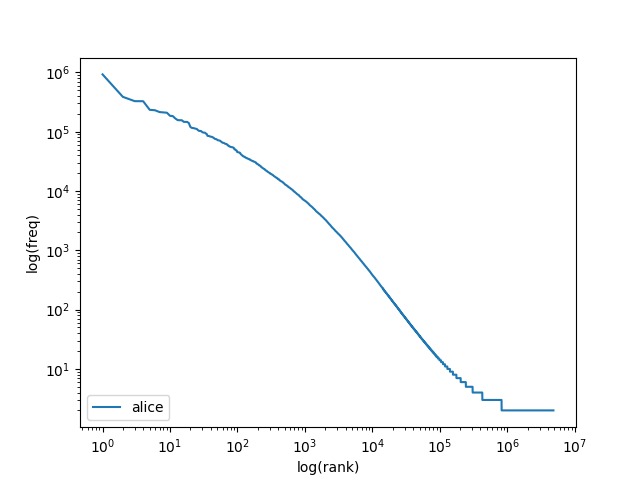
**הכי פחות שכיחים –**

**הכי שכיחים –**

* 1. הציגו את המילים הייחודיות במאגר על גרף.

גרף Zipf`sלפני ביצוע log:

****

גרף Zipf`אחרי ביצוע log:

בגלל שהקורפוס מכיל המון מילים ומופעים, בחרנו לעשות לוג על מנת להראות את התוצאות בצורה טובה יותר. כפי שניתן לראות הגרף לינארי יורד (אומנם לאט(

* 1. גודל ה: Posting –

**ללא סטמינג –** 1230000KB

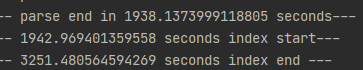
**עם סטמינג –** KB1340000

* 1. הציגו את משך הזמן שלקח למנוע לבנות את האינדקס על קבצי ה Corpus.

**פרסר** סיים לאחר בערך 32 דקות.

**אינדקס** סיים לאחר בערך 22 דקות.

**סה"כ** זמן ריצה – בערך 52 דקות.

****

* 1. עבור שאילתות מספר 8,7,4,2,1

שאילתה 1 :

|  |  |  |
| --- | --- | --- |
|  | Rank | Tweet\_id |
| Damning CNN supercut shows how Trump has been ‘wrong about every aspect of the pandemic’: President Donald Trump’s allies have been circulating an opposition research file on **Dr. Anthony Fauci**, the director of the National Institute of Allergy… | 0.939235704411646 | 1282632048759328768 |
| Indeed. Latest **published** research shows that herd immunity appears unattainable without effective vaccine. | 0.9338593328671374 | 1280945351562637317 |
| Dr. Mikovits lies about **Dr. Fauci** to hide her theft of company property from WPI, spreading misinformation about face masks. I worked for Mikovits and was identified in her first 2014 book on pages 23/42. She **published** a fraudulent Science paper: http://bit.ly/2YqT0jN | 0.9317580755057097 | 1285065590202200064 |
| The final co-author of the retracted Lancet paper, Amit N Patel, has not **published** any covid-19 papers apart from the two Surgisphere papers. He has **published** on many other topics and appears to be a productive researcher. | 0.9308333045177725 | 1290894011679608832 |
| A University of Arizona doctor just said on CNN that "#**Fauci's** a God in the infectious disease world. He wrote the book." So why was he not present or invited to today's #coronavirus #pandemic press briefing? #COVID19 #DonaldTrump | 0.930075545264682 | 1285738477208797185 |

Dr. Anthony Fauci wrote in a 2005 paper published in Virology Journal that hydroxychloroquine was effective in treating SARS.

שאילתה 2:

The seasonal flu kills more people every year in the U.S. than COVID-19 has to date.

|  |  |  |
| --- | --- | --- |
|  | Rank | Tweet\_id |
| Way more than 80K died of **Covid** and wearing a mask does not mean people don’t care about **flu** deaths. In fact I’m predicting far fewer outbreaks and deaths from **flu** this **year** if **people** all wear masks.… | 0.939235704411646 | 1284221574506151937 |
| **Flu** was shown to have over 70% of asymptomatic cases. Do you think we would lockdown and wear masks each **year** if we tested everybody for the **flu** and found out there is a massive epidemic every single year? @BAG\_OFSP\_UFSP | 0.9682994212394209 | 1281649720297832448 |
| 160 kids died from the **flu** and we don't close down every year! **Covid** is LESS deadly than the flu for those 15 and under and very similar for 21 and under. Stop using kids as your political pawns!!! | 0.9642670855013435 | 1283915280167718912 |
| “Wear a mask“ to hide your face and let child traffickers blend in and not be identified. “Wear a mask“ so **people** can riot, destroy, defame, and kill without being identified. With **seasonal flu** having the same deaths...this plandemic is more of a hoax than first thought. | 0.9626011477116306 | 1291716121054318593 |
| Are you stupid? Do I have to repeat myself how many times? I am talking about the risk for kids. Every **year** there are more deaths from **flu** FOR KIDS than from **coronavirus**. | 0.9625001121831899 | 1280966712377192450 |

שאילתה 4:

The coronavirus pandemic is a cover for a plan to implant trackable microchips and that the Microsoft co-founder Bill Gates is behind it

|  |  |  |
| --- | --- | --- |
|  | Rank | Tweet\_id |
| New article: "**Bill Gates** warns of a 'deadlier pandemic' if Covid vaccines go only to the highest bidder" has been published on Fintech Zoom - https://fintechzoom.com/fintech\_news\_hot-topic-cnbc/bill-gates-warns-of-a-deadlier-pandemic-if-covid-vaccines-go-only-to-the-highest-bidder/ @FintechZoom #Fintech #Blockchain #Crypto #Cryptocurrency #ICO | 0.8924271002433913 | 1282458952853225474 |
| Casey Welch, **Co-founder** and CEO of @AppTallo , is prepping the digital talent recruitment platform to be a resource to companies and colleges looking to rebuild, retool, and restart in the wake of the COVID-19 pandemic. @AuthorityMgzine #CHStech #talent | 0.8892374256805646 | 1285631062291480579 |
| Read my article about drivers protection during COVID19. My company Aitim Consulting Agency can help carriers to create a **plan** for handling their fleet properly and safe during the pandemic. #COVID #Trucking | 0.876424458586821 | 1283914803543789569 |
| Netflix should commission a documentary titled 'Curing the Curse' directed by David Fincher covering how few pharma companies raced frantically to develop & release Covid's vaccine with cut scenes of frontline workers, Chinese origins, trump's changing stance & **bill gates's** views | 0.8763967399485765 | 1287458680087035904 |
| . @KwantlenU instructor and founder of @NEVRCampaign , Dr. Balbir Gurm and her team, have released a new e-book on the **pandemic** of relationship violence amidst the #COVID\_\_19 pandemic. https://kpu.ca/news/2020/07/16/new-e-book-tackles-relationship-violence-pandemic | 0.8763033079467095 | 1283820196898078721 |

שאילתה 7:

Herd immunity has been reached.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Rank | | Tweet\_id |
| Thanos would implement a **herd** **immunity** policy. | 0.8869757777339253 | | 1281570352792051712 |
| It appears that some of you have contracted "compliance19". Don't worry, I have **herd** **immunity**.😂😂 | 0.8820628205767319 | | 1282430486464335874 |
| How many more recoveries before we reach **herd** **immunity**? | 0.8807987289169743 | | 1281526339313577985 |
| They’ve reached **herd immunity** | 0.8780091368944195 | | 1282410603794108417 |
| I think the strategy is clear now. Its **herd** infection hoping for **herd immunity**. It is the only play left. Biden will have a sick and divided nation to heal. I cannot believe the destruction Trump has wrought in the past 6 months alone. | | 0.8734736047865105 | 1281134812863107072 |

שאילתה 8:

Children are “almost immune from this disease.”

|  |  |  |
| --- | --- | --- |
|  | Rank | Tweet\_id |
| Because the **children** will be alone in the buildings without adults...& all the **children** are healthy, lean & do not have any compromised **immune** systems, including no asthma... & no **children** have died from this virus...& COVID19 does not scar the lungs of asymptomatic children. 🤡. | 0.9149618195476897 | 1281080670173851648 |
| Pathologist found blood clots in ‘almost every organ’ during autopsies on Covid-19 patients #usariots #usaprotest #usapolice #usaburning #usanews Autopsies on people who died of the coronavirus are helping doctors understand how the **disease** affects the … https://ift.tt/2OcrdPB | 0.899526445145981 | 1281590851542626304 |
| #Trump was right. **Children** are almost **immune** to Coronavirus. Here is why. 48% of the deaths in America occurred in Nursing homes with patients over 65 years of age. The reason for this is in how viruses attack our **immune** system and how our bodies defend themselves. | 0.8978640090623249 | 1291362821612040200 |
| Because the threat of covid is blown way out of proportion due to the fact that 1 city seemingly went out of its way to spread the **disease** among the elderly, and that **children** have almost no risk of infection let alone fatality. Just a thought. | 0.8913654749867085 | 1283847654863073280 |
| Seven months in to the coronavirus outbreak, and we still don't have a handle on why **children** are less likely to be infected, and why almost every child who does get infected has (at worst) mild symptoms. Less than 5% of confirmed infections are in those under the age of 18. | 0.8912231483660414 | 1281756084311662599 |

**אופן פעולת הדירוג שלנו הוא מציאת מילים משותפות בין שאילתה למסמך ולאחר מכן הרחבת שאילתה ע"י פונקצית דמיון של GloVe. נדרג את המסמכים באמצעות חישוב ווקטורים cosine similarity . לכן ניתן לראות בכל אחד מהשאילתות ובציוצים שחזרו שראשית קיים מילים משותפות בין השאילתה לתוצאות. שנית, בגלל הרחבת שאילה והמודל ניתן לראות כי גם מילים בעלות קשר סמנטי דומה התקבלו במסמכים- כי בעלות ווקטור דומה.**