ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

דו"ח – חלק ג'

הגישות השונות שמימשנו:

Search_engine_1 – Word2Vec

מודל זה מבוסס על רשתת נוירונים דו שכבתיות המאומנות לשחזר הקשרים לשוניים של מילים. Word2vec לוקח כקלט קורפוס גדול של טקסט ומייצר מרחב וקטורי, כאשר לכל מילה ייחודית בקורפוס מוקצה וקטור במרחב. וקטורי המילים ממוקמים במרחב הווקטורי כך שמילים החולקות הקשרים משותפים בקורפוס ממוקמות קרוב זו לזו במרחב.

אנחנו בחרנו להשתמש בארכיטקטורת skip gram, שבה המודל משתמש במילה הנוכחית כדי לחזות את החלון שמסביבה. בחרנו בשימוש שלה מכיוון שהיא עושה עבודה טובה יותר למילים נדירות.

למטרת בניית המודל השתמשנו ב- API חיצוני של gensim - ייבאנו מתוך models את Word2Vec בעזרת הפונקציות שהספריה סיפקה לנו יכלנו ליצור ולאמן את המודל שלנו.

בשלב הראשון השתמשנו בפונקציה ()Word2Vec על מנת להגדיר את הפרמטרים של המודל, את הפרמטרים הגדרנו לאחר חיפוש נרחב באינטרנט וקריאה של מדריכים ובנוסף לכך גם לאחר ניסוי וטעייה שלנו של שינויי הפרמטרים.

בשלב השני השתמשנו בפונקציה ()build_vocab על מנת לבנות את אוצר המילים מרצף משפטים ובכך לאתחל את המודל. לתוך פונקציה זו ישנה דרישה להכניס את הקורפוס שעליו נרצה לאמן את המודל (רשימה של רשימות, כך שכל רשימה מכילה את המילים במשפט). הקורפוס שעליו בחרנו לאמן את המודל הוא הקורפוס המלא מחלק א', כל המסמכים שהכנסנו למודל עברו ניקוי - כלומר עברו בפרסר שלנו, ובנוסף לכך הקטנו את כל המילים. בחרנו להוריד מרשימת המילים את היישויות, התיוגים (@) ואת ההאשטגים השארנו רק את המילים המפורקות - לדוגמא אם הופיע לנו ההאשטג הבא #DonaldTrump אז המילים שהכנסנו למודל הם donald ו-trump, זאת מכיוון שרצינו לקחת בחשבון את משמעות המילה (במידה והיתה מחוברת המשמעות שלה היתה פוחתת). החלטנו להוריד את המילים הנ"ל לאחר ניסוי וטעייה - ראינו שללא המילים האלה הגענו לתוצאות טובות יותר והמודל היה יותר מדויק, בנוסף לכך היינו מעוניינים להכניס כמה שיותר מילים רלוונטיות למודל תחת הגבלת גודל המודל ל- 150MB.

בשלב השלישי השתמשנו בפונקציה ()train על מנת לאמן את המודל.

אופן השימוש בשיטה זו בוצע במחלקת **Ranker - בזמן דירוג המסמכים**:

לאחר שקיבלנו את המסמכים הרלוונטיים מה – **Searcher**, עבור דירוג המסמכים אל מול השאילתה שהתקבלה השתמשנו בוקטורי המילים שקיבלנו מהמודל. לכל מסמך בנינו וקטור שבנוי מחיבור וקטורי המילים שנמצאות במודל, על הוקטור עשינו ממוצע וכך קיבלנו ערך למסמך, כך גם עשינו לשאילתה. ועבור כל מסמך חישבנו cos similarity עם השאילתה בעזרת הערכים שהתקבלו לנו. (הדירוג הנ"ל בוצע כתחליף לדירוג של ה- cos similarity בעזרת tf-idf).

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

<u>מפני שרצינו להגיע לזמן ריצה מיטבי, נעשו כמה החלטות בבנייה והן:</u>

- **הקטנת גודל המודל -** בעזרת הפונקציה init_sims(replace=True), פונקציה שתעזור למודל להיות הרבה יותר חסכונית בזכרון.
- **שמירת המודל בבינארי -** בעזרת הפונקציה:

save_word2vec_format('model',binary=True)

מה שעזר לצמצם משמעותית את גודל המודל ובכך להקל על זמן החזרת תשובה לשאילתה.

<u>מפני שרצינו לשפר את איכות התוצאות, נעשו כמה החלטות בבנייה והן:</u>

- **ניקוי המודל** (כפי שציינו לעיל) - השיפור המשמעותי היה בערך ה-MAP כאשר עם מודל ללא הניקוי (יישויות, @, ו-#) הגענו לערך של 0.752 ועם הניקוי הגענו לערך של 0.765.

- **דירוג המסמכים -** ראינו שלעומת הדירוג של ה- cos similarity בעזרת tf-idf השיפור היה משמעותי בערך ה-MAP. כאשר בעזרת דירוג של cos similarity בעזרת tf-idf קיבלנו ערך של 0.735 ועם הדירוג cos similarity בעזרת וקטורי המילים מהמודל קיבלנו ערך של 0.765.

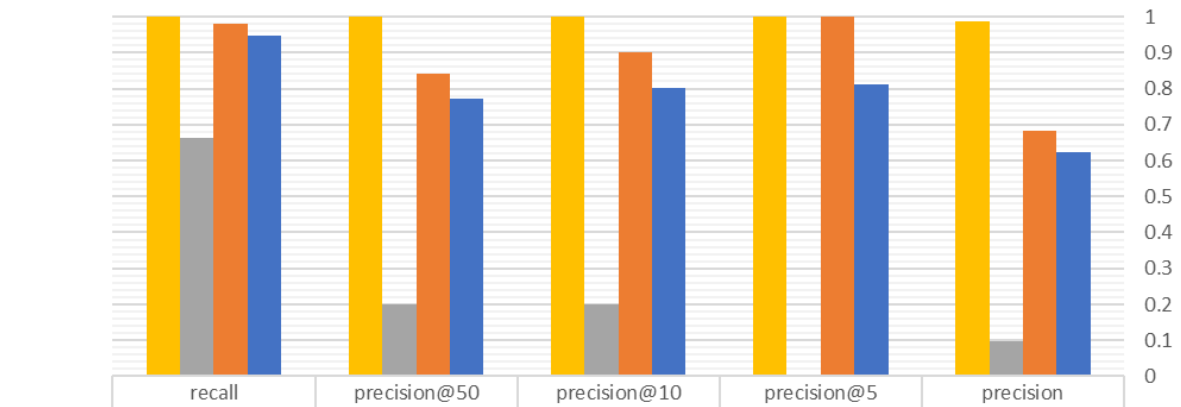
- **החלפת מילים מהשאילתה** - כאשר נקבל שאילתה נבדוק אם כל המילים מופיעות באינדקסר שבנינו - במידה והמילה לא נמצאת נרצה לחפש מילה הכי קרובה שכן נמצאת באינדקסר וזאת על מנת שנוכל לקבל כמה שיותר מסמכים שקרובים לקורפוס שאותו נקבל. כאשר קיבלנו בשאילתה מילה שלא נמצאת באינדקסר, נבדוק אם המילה נמצאת במודל ונחפש לה את המילה הכי קרובה מהמודל שנמצאת באינדקסר. במידה ומצאנו נחליף את המילה הזו.

<u>כמה נקודות למחשבה שעלו לנו בעת שימוש בשיטה שכזו, היינו מעוניינים לנצל את הפוטנציאל כמה שיותר, אך לא ראינו שיפור משמעותי בעזרתן:</u>

- **הרחבת השאילתה** - נעבור על כל המילים בשאילתה ונבדוק איזה מילים נמצאות במודל שלנו, נכניס את כל המילים מהשאילתה שנמצאות במודל לפונקציה most_similar (פונקציה שנמצאת באובייקט של המודל שבנינו), נקבל את המילה שהכי קרובה לרשימת המילים וגם שהמילה נמצאת באינדקסר ונוסיף אותה לשאילתה - עשינו זאת במטרה לנסות לקבל מסמכים נוספים שדומים בקונספט שלהם למילים בשאילתה.
- **צימצום השאילתה** - נעבור על כל המילים בשאילתה ונבדוק איזה מילים נמצאות במודל שלנו, נכניס את כל המילים מהשאילתה שנמצאות במודל לפונקציה doesnt_match (פונקציה שנמצאת באובייקט של המודל שבנינו), נקבל את המילה שהכי לא מתאימה מרשימת המילים שהכנסנו ונוריד אותה מהשאילתה - עשינו זאת במטרה לנסות להוריד מסמכים שהמילה שיצאה כלא מתאימה נמצאת בהן (כדי להוריד מסמכים שלא באמת רלוונטים אך חזרו כרלוונטים).

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

תוצאות המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.947103463	0.772571429	0.802857143	0.811428571	0.623498048
median	0.979166667	0.84	0.9	1	0.681818182
min	0.663157895	0.2	0.2	0	0.097165992
max	1	1	1	1	0.985347985

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט השאילתות:

על פי השיטה **Word2Vec** נשים לב שככל שהמסמך דורג במקום יותר גבוה זה אומר שהמילים שנמצאות בשאילתה והמילים שנמצאות במסמך חולקות הקשרים משותפים רבים יותר בקורפוס ולכן וקטור המסמך יהיה קרוב לוקטור השאילתה ולכן ממוקם בדירוג גבוה מבין המסמכים.

שאילתה מספר 1 -

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

fauci paper hydroxychloroquine sars (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבשלושת המסמכים הראשונים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים חולקות הקשרים משותפים רבים עם המילים בשאילתה. בנוסף לכך, נראה שעבור המסמך הרביעי והחמישי אמנם הקשר מאוד דומה אך לא חזק כמו בשלושה הראשונים.

.1

RT @StoneSculptorJN: Fauci's own NIH published a paper showing that Chloroquine is a "potent inhibitor of SARS" (but there is little money**…**

.2

RT @USAMomUtah: @CindyProUSA Yes. Dr Fauci published a paper in 2005 praising the use of hydrochloroquine in the originals SARS COVID outbr**…**

.3

Dr Fauci wrote medical paper on

Hydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he down plays it now

.4

@EvilDave_NXT @MichaelCoudrey U R LYING !!! @drdavidsamadi @zev_dr @raoult_didierOf DR ANTHONY FAUCI Said himself in 2005 that chloroquine, hydroxychloroquine are definitely an effective therapeutic and prophylactic against corona viruses check it out @realDonaldTrump @maddow @CNN @raoult_didierOf @CNN**!!!**

.5

RT @mitchellvii: Here's #DrFraudFauci in 2005 calling #Hydroxychloroquine a "wonder drug" in defeating SARS virus. https://t.co/XOSKpI1ZpX

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שהמסמך הראשון הוא בעל הקשר החזק ביותר עבור המילים שמופיעות גם במסמך וגם בשאילתה, הן חולקות הכי הרבה הקשרים משותפים. נראה ששאר המסמכים גם יש הקשרים משותפים אך פחות מאשר המסמך הראשון.

.1

I guess COVID-19 kills the flu?

.2

@OHaraTony @QDROP8 Covid-19 IS the seasonal flu.

.3

Is COVID-19 deadlier than the flu? https://t.co/nbWx5HOqhP

.4

@ConservBlue2020 @RyanAFournier No, its not....the flu kills more people than Covid....get a grip.

.5

@itvnews Would this be the seasonal flu? Renamed as covid19?

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נראה שחזרו מסמכים שונים אך עם טקסט דומה (מכיוון שזה **RT**), המילים שמופיעות במסמך חולקות הקשרים משותפים עם המילים שנמצאות בשאילתה.

.1

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.2

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.3

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.4

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.5

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבארבעת המסמכים הראשונים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים מופיעות בהקשרים משותפים רבים עם המילים בשאילתה. בנוסף לכך, נראה שעבור המסמך החמישי אמנם הקשר מאוד דומה אך לא חזק כמו באחרים.

.1

@yashar They probably reached herd immunity....

.2

Sure did, and they have reached herd immunity

.3

@clarkle119 @benshapiro Northeast has reached herd immunity.

.4

@Monideepa62 We have reached herd immunity to injustice.

.5

@BeckiMeister @philipoconnor How is this not herd immunity? https://t.co/54oIh9h2Cb

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבכל המסמכים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים מופיעות בהקשרים משותפים רבים עם המילים בשאילתה.

.1

"Children almost immune from COVID"!

.2

No, children are NOT immune to Covid-19.

.3

@Reuters Children are almost immune to the coronavirus.

That's a fact.

.4

Children are Almost Immune from COVID-19

.5

@RaheemKassam @JackPosobiec Was a clip from Fox saying children are almost immune from #coronavirus

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט תוצאות המדדים של כל השאילתות:

search engine 1					
recall	precision @50	precision @10	precision @5	precision	Query num
0.984293	0.86	1	1	0.696296	1
1	0.8	0.6	0.6	0.758242	2
0.995726	0.9	1	1	0.862963	3
0.993976	0.96	1	1	0.627376	4
0.663158	0.3	0.5	0.4	0.372781	5
0.983607	0.42	0.2	0	0.43956	6
0.831169	0.98	1	1	0.845815	7
0.996296	0.98	0.9	0.8	0.985348	8
1	0.74	0.6	0.2	0.778182	9
0.995671	0.86	0.9	0.8	0.884615	10
1	0.98	1	1	0.795455	11
0.96	0.2	0.7	1	0.097166	12
0.928571	0.54	0.9	1	0.193069	13
0.979167	0.88	0.8	0.6	0.917969	14
0.941176	0.96	0.9	1	0.806723	15
0.951613	0.86	0.7	1	0.797297	16
0.758865	0.9	0.9	0.8	0.84252	17
0.984615	0.8	0.7	1	0.747082	18
0.927536	0.8	0.7	0.6	0.528926	19
0.954545	0.8	1	1	0.276316	20
0.95614	0.84	0.9	1	0.465812	21
0.977528	0.54	0.4	0.4	0.332061	22
1	0.76	0.6	0.8	0.621514	23
0.990991	0.78	0.7	0.6	0.578947	24
0.946429	0.5	0.6	0.6	0.215447	25
0.982143	0.96	1	1	0.681818	26
0.994681	0.98	1	1	0.760163	27
0.80597	0.72	0.9	1	0.610169	28
0.811688	0.92	1	1	0.589623	29
0.951923	0.68	0.9	1	0.518325	30
0.938144	0.84	0.9	1	0.758333	31
1	0.32	0.3	0.2	0.251938	32
0.995968	1	1	1	0.94636	33
0.989418	0.94	1	1	0.730469	34
0.977612	0.74	0.9	1	0.507752	35
0.947103	0.772571	0.802857	0.811429	0.623498	average
0.979167	0.84	0.9	1	0.681818	median
0.663158	0.2	0.2	0	0.097166	min
1	1	1	1	0.985348	max
0.765074911					MAP

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

Search_engine_2 – Global Method

השתמשנו בשיטה Global Method בתצורת Query Expansion. השיטה מבוססת על מטריצת אסוציאציות בין המילים באינדקס. יחושבו יחסים בין המילים השונות לפי הופעתם במסמכים, ובסוף בניית האינדקס (כאשר נעבור על כל המסמכים) - נבצע נרמול לערכי הדמיון כך שהערכים של הדמיון [0,1]. בעקבות כך, נדע שמילים שערך הדמיון שלהם הינו 1 הן מילים שתמיד הופיעו ביחד, אותו מספר פעמים בדיוק. באשר לערך דמיון 0 - אלו מילים שלא הופיעו ביחד בכלל. לכן, שימוש בערכים הנ"ל בשלב אחזור השאילתה לשם הוספת מילים לשאילתה (Query Expansion) הינו צעד שישפר את יכולות האחזור של המנוע שלנו (ככל שערך הדמיון גבוה יותר) כי יכוון את ה-Searcher למסמכים יותר רלוונטיים. לשם מימוש השיטה לא היה צורך ב-API ולא בקוד פתוח.

מפני שרצינו להגיע לזמן ריצה מיטבי, נעשו כמה החלטות בבנייה והן:

- **צמצום מטריצת הדמיון, על ידי שמירת הערכים הנחוצים בלבד**: נשמור במילון עבור כל Tuple של המילים ((2w1,w) = כאשר תמיד 2w1<w לקסיקוגרפית לשם חד-חד-ערכיות בשמירה הערכים וחיסכון בבלבול), כך ש"חתכנו" בחצי כי לא שמרנו עבור צמד מילים את - (w1,w2) .בנוסף, כמו שנאמר בפסקה הקודמת - מיותר לשמור ערכים שלא יועילו לחישוב ולכן לא שמרנו את צמדי המילים אשר לא מופיעים יחדיו באותו מסמך בכלל (ערך ה-Sij שלהן יהיה 0) - ובכך "חתכנו" שוב, ובנינו כביכול מטריצה דלילה במבנה נתונים מילון.

- **צמצום זמן ריצה באופן ישיר, על ידי גישה מהירה יותר לערך**: בעת חישוב Sij (הערך המנורמל של דמיון בין צמד מילים (i,j), צריך את Cii,Cjj,Cij - שהם ערכי דמיון לא מנורמלים בין אותה מילה, לעצמה, ואת היחס בין שתי המילים השונות. לכן, לשם צמצום זמן ריצה שמרנו בקובץ האינדקס ההופכי שדה עבור יחס המילה עם עצמה (Cii). כך בשעת חישוב - ישנו מעבר רק על מילון יחסי המילים (כי נזניח גישה ב-O(1 למפתח במילון הקובץ ההופכי, נזניח בחישוב זמן ריצה התנגשויות ב-Hash).

- ב**שלב הרחבת השאילתה, הגדרנו סף לסיום מעבר על צמדי המילים וערכיהן**: על ידי מיון המילון על פי Sij (בסדר יורד), ועצירה בתנאים מאוד ספציפיים ומדודים (בפסקה הבאה).

בנוסף, ניסינו לשפר את איכות האחזור על ידי הסרת מילים שמופיעות פעם אחת בקורפוס:

כשניתחנו את ערכי מטריצת האסוציאציות, ראינו שישנם הרבה צמדי מילים שמופיעות פעם אחת בכל הקורפוס, ולהן Sij אשר שווה ל-1 (המקסימום). לכן, הסקנו שאיכות האחזור עלולה להשתפר עקב הסרת מילים אלו. עקב העובדה שמעשה זה לא שיפר לנו את התוצאות, 2ויתרנו עליו.

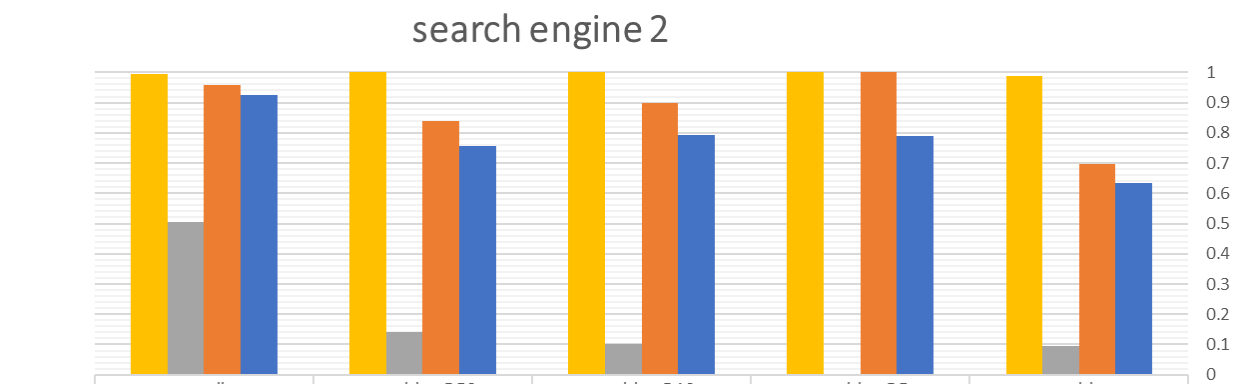
מכיוון שרצינו להגיע לאיכות תוצאות מיטבית, נעשו כמה החלטות בשלב "הרחבת השאילתה" והן קורות בשלב בו נעבור על כל המילים בשאילתה עבור כל צמד מילים במילון.

- **נבצע הרחבה לשאילתה עם מילה, אם ערך הדמיון גדול או שווה ל- 0.2** : לאחר ניסיונות, בדיקות מעמיקות והשוואת מדדים שהגדרנו כרלוונטיים לטיב האחזור - הגענו למסקנה שמילים שערך הדמיון קטן מ-0.2, אינן רלוונטיות ופוגעות באחזור.

- **נבצע הרחבה לשאילתה עם מילה אחת, שתי מילים או בכלל לא :** ראינו לנכון על פי המדדים שהוגדרו מראש לא להוסיף יותר משתי מילים לשאילתה (לא נרצה לאבד את יתרון השיטה על ידי שאילתה גדולה מדי).

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

תוצאות המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.92377202	0.756571429	0.791428571	0.788571429	0.63431201
median	0.957894737	0.84	0.9	1	0.697761194
min	0.506493506	0.14	0.1	0	0.094827586
max	0.995169082	1	1	1	0.98880597

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

<u>פירוט השאילתות:</u>

על פי השיטה **Global Method** ביצענו **query expansion** – המילים שמסומנות באדום הן המילים שהתווספו לשאילתה לאחר ההרחבה. מכיוון שכאן הדירוג מבוצע בעזרת **cos similarity** בעזרת **tf-idf** ישנה חשיבות לכמות המילים המשתופות עם השאילתה.

<u>שאילתה מספר 1 -</u>

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

fauci paper hydroxychloroquine sars (keywords)

<u>הציוצים שהוחזרו:</u>

.1

RT @StoneSculptorJN: Fauci's own NIH published a paper showing that Chloroquine is a "potent inhibitor of SARS" (but there is little money)

על פי השיטה Global Method, חזרו 3 מתוך 4 מילים מ-keywords. התווספו published , cov. ניתן לראות שהמילה published מופיעה בשאילתה המלאה וגם בציוץ שחזר.

.2

Dr Fauci wrote medical paper on Hydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he downplays it now

על פי השיטה Global Method, חזרו 4 מתוך 4 מילים מ-keywords. התווספו published , cov.

.3

@ScottAdamsSays Also, why is HCQ a "Chloroquine is a potent inhibitor of SARS coronavirus infection and spread," but not SARS-CoV-2? Paper is from 2005
https://t.co/QqSWDo7Erg

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו published , cov. ניתן לראות שהציוץ שחזר יש את המילה cov.

If hydroxychloroquine, azithromycin & zinc don't work why do they care about us having access to them?????

על פי השיטה Global Method, חזרו 1 מתוך 4 מילים מ-keywords. התווספו published , cov.

.4

If hydroxychloroquine, azithromycin & zinc doesn't work why do they care about us having access to it?

על פי השיטה Global Method, חזרו 1 מתוך 4 מילים מ-keywords. התווספו published , cov.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

.1

@MarkjgloverJ @NHSuk @Tesco @asda @sainsburys Covid-19 is not the flu.

Flu has a vaccine, covid-19 does not

Flu has a fast incubation period.

Flu kills 10k a year in the uk, covid has killed 45k so far this year.

You can have covid-19 and not know it, but you can spread it, wearing a face covering stops that.

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו every ,19.
ניתן לראות שהציוץ שחזר יש את המילה 19.

.2

@FaheemYounus The flu kills hundreds of thousands of people each year, and for most people, Covid is much less dangerous than the flu.

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו every ,19.

.3

my family has started the “the flu kills more people each year than covid” argument with me when actually no? it kills tens of thousands of people in an entire year when covid has killed over 150,000 people in 6 months even with the lockdowns so like? ??????????

על פי השיטה Global Method, חזרו 3 מתוך 4 מילים מ-keywords. התווספו every ,19.

.4

The Seasonal Flu kills approximately 650000 people world wide A YEAR

COVID19 has killed 706000 people so far.(Take note that is extremely inflated)

The kick is that the Flu has a vaccine and this many people die. EVERY YEAR.

Why are we worked up over COVID but not the flu?

על פי השיטה Global Method, חזרו 3 מתוך 4 מילים מ-keywords. התווספו every ,19.
ניתן לראות שהציוץ שחזר יש את המילה every,19.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

@fatbonkrips @smith_tarina @GovInslee So you are going to stay home all flu season every year? Because the flu kills more people every year as COVID is going to this year (except in a few areas like NYC maybe)

על פי השיטה Global Method, חזרו 3 מתוך 4 מילים מ-keywords. התווספו every ,19.
ניתן לראות שהציוץ שחזר יש את המילה every,19.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

.1

Vaccines to implant tracking devices? Bill Gates goes public https://t.co/R51SAQ3ueD

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords. התווספו bill,devices.
ניתן לראות שהציוץ שחזר יש את המילה bill,devices.

.2

Vaccines to implant tracking devices? Bill Gates goes public https://t.co/KHnke7ATjA

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords. התווספו bill,devices.
ניתן לראות שהציוץ שחזר יש את המילה bill,devices.

.3

Vaccines to implant tracking devices? Bill Gates goes public https://t.co/LXmdbsERRt

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords. התווספו bill,devices
ניתן לראות שהציוץ שחזר יש את המילה bill,devices.

.4

Vaccines to implant tracking devices? Bill Gates goes public https://t.co/gLYwByiqCz IS CRAZY!!

.5 על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords. התווספו bill,devices
ניתן לראות שהציוץ שחזר יש את המילה bill,devices.

.6

Vaccines to implant tracking devices? Bill Gates goes public https://t.co/BPPnJamJd7 fuk gates !

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords. התווספו bill,devices
ניתן לראות שהציוץ שחזר יש את המילה bill,devices.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

.1

What the Doctor says

There is no herd immunity without a vaccine. There is no herd immunity without a vaccine. There is no herd immunity without a vaccine.

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords.

.2

Herd immunity reached long b4 a vaccine available.

על פי השיטה Global Method, חזרו 3 מתוך 3 מילים מ-keywords.

.3

@ellymelly Have we reached herd immunity for the common cold?

על פי השיטה Global Method, חזרו 3 מתוך 3 מילים מ-keywords.

.4

Why herd immunity to COVID-19 is reached much earlier than thought https://t.co/Yp5Jl7nr45 ???

על פי השיטה Global Method, חזרו 3 מתוך 3 מילים מ-keywords.

.5

@PandemicPanda10 @IngrahamAngle As well as immunity from antibodies, there is natural immunity, there is immunity from T-cells, and there is cross-immunity from other coronaviruses. Germany, France, Sweden, and New York have reached herd immunity. The rest of the US is only a few weeks behind. Accept reality.

על פי השיטה Global Method, חזרו 2 מתוך 3 מילים מ-keywords.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

הציוצים שהוחזרו:

.1

Video: Fact check: Children are not immune, or almost immune, from the virus. https://t.co/tiH2W99Lfa

על פי השיטה Global Method, חזרו 3 מתוך 4 מילים מ-keywords. התווספו almost. ניתן לראות שהציוץ שחזר יש את המילה almost.

.2

The president just said “children are almost immune” to covid

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו almost. ניתן לראות שהציוץ שחזר יש את המילה almost.

.3

trump, “children are almost immune to Covid.” Sure.....

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו almost. ניתן לראות שהציוץ שחזר יש את המילה almost.

.4

RT @therecount: @Yamiche Fact check: Children are not immune, or almost immune, from the virus. https://t.co/C77x1tSKrD

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו almost. ניתן לראות שהציוץ שחזר יש את המילה almost.

.5

RT @therecount: @Yamiche Fact check: Children are not immune, or almost immune, from the virus. https://t.co/C77x1tSKrD

על פי השיטה Global Method, חזרו 2 מתוך 4 מילים מ-keywords. התווספו almost. ניתן לראות שהציוץ שחזר יש את המילה almost.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט תוצאות המדדים של כל השאילתות:

search engine 2					
recall	precision @50	precision@ 10	precision @5	precision	Query num
0.979058	0.84	0.7	1	0.697761194	1
0.995169	0.8	0.8	0.8	0.768656716	2
0.970085	1	1	1	0.879844961	3
0.987952	0.96	0.9	1	0.625954198	4
0.957895	0.36	0.3	0	0.356862745	5
0.959016	0.44	0.5	0.4	0.436567164	6
0.506494	0.94	0.9	0.8	0.866666667	7
0.981481	1	1	1	0.98880597	8
0.990654	0.72	0.6	0.8	0.794007491	9
0.974026	0.92	0.8	0.8	0.9	10
0.990476	0.96	1	1	0.79389313	11
0.88	0.2	0.7	1	0.094827586	12
0.928571	0.64	1	1	0.193069307	13
0.991667	0.94	1	1	0.911877395	14
0.931373	0.84	0.6	0.8	0.798319328	15
0.951613	0.96	1	1	0.793721973	16
0.673759	0.92	0.9	1	0.904761905	17
0.948718	0.9	0.8	0.6	0.74	18
0.927536	0.66	0.7	1	0.533333333	19
0.909091	0.54	0.5	0.4	0.289855072	20
0.95614	0.72	0.9	1	0.465811966	21
0.977528	0.5	0.5	0.4	0.334615385	22
0.987179	0.88	0.7	0.4	0.633744856	23
0.981982	0.84	0.9	0.8	0.579787234	24
0.803571	0.14	0.1	0	0.194805195	25
0.97619	0.94	1	1	0.686192469	26
0.978723	0.98	1	1	0.796536797	27
0.738806	0.84	1	1	0.61875	28
0.902597	0.92	1	1	0.579166667	29
0.836538	0.9	1	1	0.763157895	30
0.938144	0.8	0.9	1	0.758333333	31
0.984615	0.18	0.3	0	0.25	32
0.919355	0.96	1	1	0.94214876	33
0.968254	0.78	0.9	1	0.729083665	34
0.947761	0.56	0.8	0.6	0.5	35
0.923772	0.756571	0.7914286	0.788571	0.63431201	average
0.957895	0.84	0.9	1	0.697761194	median
0.506494	0.14	0.1	0	0.094827586	min
0.995169	1	1	1	0.98880597	max
0.745140537					MAP

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

Search_engine_3 – Spelling Correction

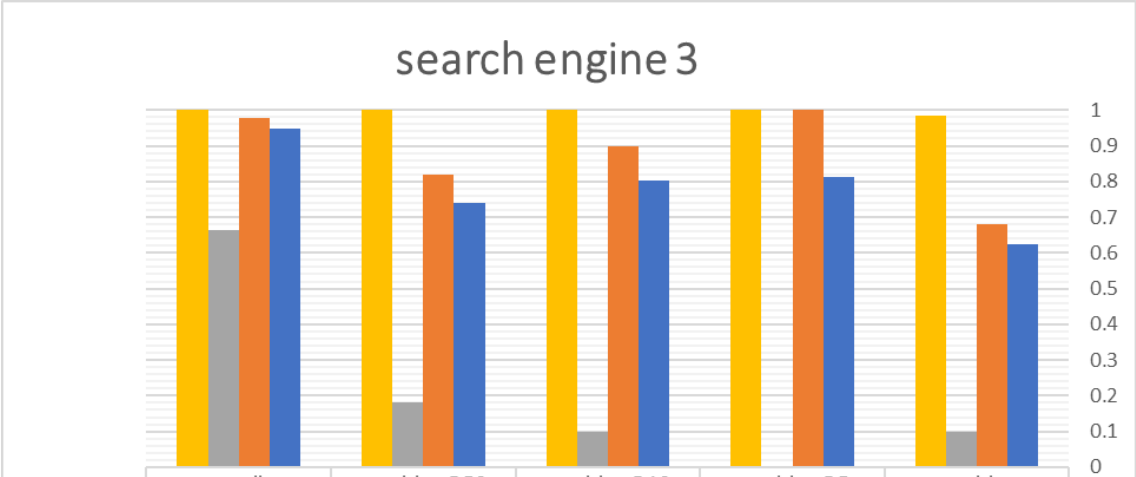
השתמשנו בשיטה Spelling Correction בתצורת Query Replacement - כלומר, החלפת המילה "השגויה" תחת תנאים מסוימים. השיטה משתמשת בקוד פתוח ושמו Pyspellchecker, ומחזירה רשימה. הרשימה מחזיקה את המילים שיש סיכוי שהינם "התיקון לשגיאה".

לא הייתה דרך ממשית להיטיב עם זמן הריצה, אז החלטנו להשתמש במתודה correction שתחזיר לנו את המילה שבהסתברות הכי גבוהה הינה המילה התקינה.

לשם שיפור איכות האחזור - החלטנו לבצע Query Replacement אך ורק עם מילים העומדות בתנאים הבאים:

- ראשית, אם המילה "השגויה" מהשאילתה נמצאת במילון הקובץ ההופכי - לא נחליף אותה.
- אם המילה "התקינה" (ממתודה **correction**) לא נמצאת במילון הקובץ ההופכי - לא נחליף איתה.
- אם המילה "התקינה", היא המילה ה"השגויה" - לא נחליף אותה.
- אם המילה "התקינה", נמצאת בשאילתה כבר - לא נחליף איתה.

תוצאות המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.946663903	0.74	0.802857143	0.811428571	0.623414884
median	0.979166667	0.82	0.9	1	0.681818182
min	0.663157895	0.18	0.1	0	0.097165992
max	1	1	1	1	0.985347985

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט השאליתות:

מכייוון שכאן הדירוג מבוצע בעזרת **cos similarity** בעזרת **tf-idf** ישנה חשיבות לכמות המילים המשתופות עם השאילתה. התיקון של השאילתה בוצע רק כאשר המילים שקיבלנו בשאילתה אינן מופיעות בקובץ ההופכי. מכיוון שכל השאילתות הוקלדו ללא שגיאות ובנוסף לכך גם הופיעו בקובץ ההופכי לא היה צורך בתיקונן ולכן לא ניתן לראות את תרומתה של השיטה **Spelling Correction**.

שאילתה מספר 1 -

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

fauci paper hydroxychloroquine sars (keywords)

הציוצים שהוחזרו:

.1

Dr Fauci wrote medical paper on \nHydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he down plays it now

על פי השיטה Spelling Correction, חזרו 4 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.2

@ScottAdamsSays Also, why is HCQ a "Chloroquine is a potent inhibitor of SARS coronavirus infection and spread," but not SARS-CoV-2? Paper is from 2005

https://t.co/QqSWDo7Erg

על פי השיטה Spelling Correction, חזרו 2 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.3

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…

על פי השיטה Spelling Correction, חזרו 1 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.4

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…

על פי השיטה Spelling Correction, חזרו 1 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.5

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…

על פי השיטה Spelling Correction, חזרו 1 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

.1

I guess COVID-19 kills the flu?

על פי השיטה Spelling Correction, חזרו 3 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.2

Is COVID-19 deadlier than the flu? https://t.co/nbWx5HOqhP

על פי השיטה Spelling Correction, חזרו 2 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.3

@FaheemYounus The flu kills hundreds of thousands of people each year, and for most people, Covid is much less dangerous than the flu.

על פי השיטה Spelling Correction, חזרו 3 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.4

@rob_miller12345 Kills less people that the flu? Based on what? You do realise we don't lockdown for the flu?

על פי השיטה Spelling Correction, חזרו 3 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.5

@KeanuTrades @Investingcom Yet flu kills more during the flu season in those age ranges than COVID-19 will year round. Oh, and there is a flu vaccine to boot. Hope you are wearing that mask during flu season.

על פי השיטה Spelling Correction, חזרו 3 מתוך 4 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

.1

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.2

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.3

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.4

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.5

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

.1

Sure did, and they have reached herd immunity

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.2

About herd immunity #covid19

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.3

Herd immunity reached long b4 a vaccine available.

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.4

@ellymelly Have we reached herd immunity for the common cold?

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.5

Why herd immunity to COVID-19 is reached much earlier than thought https://t.co/Yp5Jl7nr45 ???

על פי השיטה Spelling Correction, חזרו 3 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

הציוצים שהוחזרו:

.1

No, children are NOT immune to Covid-19.

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.2

"Children almost immune from COVID!"

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.3

Children are Almost Immune from COVID-19

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.4

Video: Fact check: Children are not immune, or almost immune, from the virus. https://t.co/tiH2W99Lfa

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

.5

Children are nearly immune from Covid disease!

על פי השיטה Spelling Correction, חזרו 2 מתוך 3 מילים מ-keywords. לא הוחלפו מילים בשאילתה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט תוצאות המדדים של כל השאילתות:

search engine 3					
recall	precision @50	precision @10	precision @5	precision	Query num
0.984293	0.82	0.6	0.8	0.696296	1
1	0.76	0.6	0.4	0.758242	2
0.995726	0.98	0.9	0.8	0.862963	3
0.993976	0.96	0.9	0.8	0.627376	4
0.663158	0.26	0.1	0.2	0.372781	5
0.983607	0.44	0.5	0.4	0.43956	6
0.831169	0.9	0.9	1	0.845815	7
0.996296	0.98	0.9	0.8	0.985348	8
1	0.68	0.7	0.4	0.778182	9
0.995671	0.9	0.8	0.6	0.884615	10
1	0.98	1	1	0.795455	11
0.96	0.22	0.7	1	0.097166	12
0.928571	0.64	1	1	0.193069	13
0.979167	0.9	0.9	1	0.917969	14
0.941176	0.8	0.7	0.6	0.806723	15
0.951613	0.96	1	1	0.797297	16
0.758865	0.88	1	1	0.84252	17
0.984615	0.86	1	1	0.747082	18
0.927536	0.74	0.7	1	0.528926	19
0.954545	0.52	0.7	0.8	0.276316	20
0.95614	0.72	0.9	1	0.465812	21
0.977528	0.4	0.6	0.8	0.332061	22
1	0.84	0.6	1	0.621514	23
0.990991	0.72	0.8	0.6	0.578947	24
0.946429	0.34	0.8	0.8	0.215447	25
0.982143	0.88	0.8	0.8	0.681818	26
0.994681	0.98	1	1	0.760163	27
0.80597	0.76	1	1	0.610169	28
0.811688	0.9	1	1	0.589623	29
0.951923	0.76	1	1	0.518325	30
0.938144	0.82	0.9	1	0.758333	31
0.984615	0.18	0.3	0	0.249027	32
0.995968	1	1	1	0.94636	33
0.989418	0.84	1	1	0.730469	34
0.977612	0.58	0.8	0.8	0.507752	35
0.946664	0.74	0.802857	0.811429	0.623415	average
0.979167	0.82	0.9	1	0.681818	median
0.663158	0.18	0.1	0	0.097166	min
1	1	1	1	0.985348	max
0.735965266					MAP

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

Search_engine_4 – Thesaurus

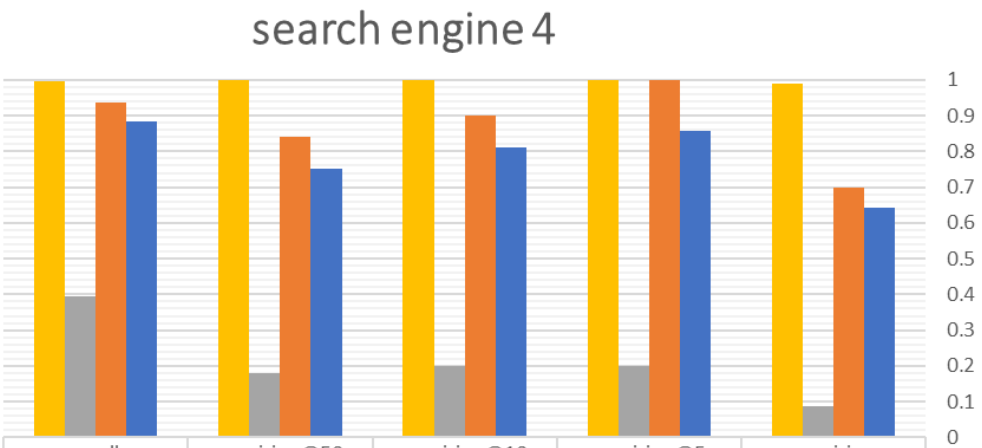
השתמשנו בשיטה Thesaurus בתצורת Query Expansion. השיטה משתמשת בקוד פתוח ושמו Lin Thesaurus, אשר שייך ל-NLTK. השיטה מחזירה מילון עם מילים נרדפות לכל מילה מהשאילתה, עם ציון מנורמל המציין כמה המילה קרובה במשמעותה למילה מהשאילתה.

לשם מיטוב איכות תוצאות האחזור, החלטנו לבצע Query Expansion אך ורק עם מילים העומדות בתנאים הבאים:

- המילה הנרדפת נמצאת במילון הקובץ ההופכי
- המילה הנרדפת לא נמצאת בשאילתה
- לאחר בדיקה, תבוצע הוספה ראשונית לרשימה, שממנה נדרג ונסנן את ה- 2 הנרדפות עם הציון הכי גבוה כך:
  - **מונה השאילתה לא יעבור את 3 :** מונה סופר כמות הוספות עבור כלל השאילתה
  - **מונה המילה מהשאילתה לא יעבור את 1 :** המונה סופר את כמות המילים שהוספנו באיטרציה על מילה ספציפית, ומתאפס עבור כל מילה מחדש.

נקודה למחשבה - לשם שיפור איכות התוצאות, שילבנו את השיטה עם שיטת Global Method בעזרת החלפת דירוג המילים הנרדפות - בדירוג ה-Sij לשם הוספת המילים הללו. אך השיטה לא הניבה שיפורים משמעותיים.

פירוט המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.882575046	0.752571429	0.811428571	0.857142857	0.643777291
median	0.936936937	0.84	0.9	1	0.700374532
min	0.393939394	0.18	0.2	0.2	0.085972851
max	0.995169082	1	1	1	0.98880597

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט השאילתות:

על פי השיטה **Thesaurus** ביצענו **query expansion** – המילים שמסומנות באדום הן המילים שהתווספו לשאילתה לאחר ההרחבה. מכיוון שכאן הדירוג מבוצע בעזרת **cos similarity** בעזרת **tf-idf** ישנה חשיבות לכמות המילים המשתופות עם השאילתה.

שאילתה מספר 1 -

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

fauci paper hydroxychloroquine sars (keywords)

הציוצים שהוחזרו:

.1

Dr Fauci wrote medical paper on

Hydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he down plays it now

על פי השיטה Thesaurus, חזרו 4 מתוך 4 מילים מ-keywords. התווספו book,document.

.2

@ScottAdamsSays Also, why is HCQ a "Chloroquine is a potent inhibitor of SARS coronavirus infection and spread," but not SARS-CoV-2? Paper is from 2005

https://t.co/QqSWDo7Erg

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו book,document.

.3

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be

על פי השיטה Thesaurus, חזרו 1 מתוך 4 מילים מ-keywords. התווספו book,document.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.4

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be

על פי השיטה Thesaurus, חזרו 1 מתוך 4 מילים מ-keywords. התווספו book,document.

.5

RT @RichHiggins_DC: Fauci wasn't "wrong."

Fauci was "lying."

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be

על פי השיטה Thesaurus, חזרו 1 מתוך 4 מילים מ-keywords. התווספו book,document.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

.1

There's been more people dying of flu and pneumonia in this year than this stupid Covid 19

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו pneumonia,influenza. ניתן לראות שחזרה גם המילה pneumonia.

.2

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו pneumonia,influenza. ניתן לראות שחזרה גם המילה pneumonia.

.3

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו pneumonia,influenza. ניתן לראות שחזרה גם המילה pneumonia.

.4

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו pneumonia,influenza. ניתן לראות שחזרה גם המילה pneumonia.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו pneumonia,influenza. ניתן לראות שחזרה גם המילה pneumonia.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

.1

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו transplant,vaccine. ניתן לראות שהמילה vaccine מופיעה

.2

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו transplant,vaccine. ניתן לראות שהמילה vaccine מופיעה

.3

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו transplant,vaccine. ניתן לראות שהמילה vaccine מופיעה

.4

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו transplant,vaccine. ניתן לראות שהמילה vaccine מופיעה

.5

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו transplant,vaccine. ניתן לראות שהמילה vaccine מופיעה

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

.1

Herd immunity reached long b4 a vaccine available.

על פי השיטה Thesaurus, חזרו 3 מתוך 3 מילים מ-keywords. התווספו exemption, protection, passed.

.2

Why herd immunity to COVID-19 is reached much earlier than thought https://t.co/Yp5Jl7nr45???

על פי השיטה Thesaurus, חזרו 3 מתוך 3 מילים מ-keywords. התווספו exemption, protection, passed.

.3

RT @js100js100: @21WIRE @JS49 And on herd immunity I can't see any other explanation for this, than herd immunity has been reached

על פי השיטה Thesaurus, חזרו 3 מתוך 3 מילים מ-keywords. התווספו exemption, protection, passed.

.4

RT @js100js100: @21WIRE @JS49 And on herd immunity I can't see any other explanation for this, than herd immunity has been reached

על פי השיטה Thesaurus, חזרו 3 מתוך 3 מילים מ-keywords. התווספו exemption, protection, passed.

.5

@daniellevitt22 @InProportion2 So has Delhi reached herd immunity? Does herd immunity mean zero cases? @PrasPro

https://t.co/BznRjRcT17

על פי השיטה Thesaurus, חזרו 3 מתוך 3 מילים מ-keywords. התווספו exemption, protection, passed.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

הציוצים שהוחזרו:

.1

Video: Fact check: Children are not immune, or almost immune, from the virus.
https://t.co/tiH2W99Lfa

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו vulnerable , susceptible.

.2

The president just said “children are almost immune” to covid

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו vulnerable , susceptible.

.3

So Trump says children are "almost immune from this disease" and that gets taken down for claiming someone is immune from the coronavirus, which it does not say.

He clearly says they are LESS susceptible.

Why pretend you don't get that?

Why should I listen to your analysis?

על פי השיטה Thesaurus, חזרו 3 מתוך 4 מילים מ-keywords. התווספו vulnerable , susceptible. ניתן לראות שהמילה susceptible נוספה.

.4

trump, “children are almost immune to Covid.”

Sure.....

על פי השיטה Thesaurus, חזרו 2 מתוך 4 מילים מ-keywords. התווספו vulnerable , susceptible. ניתן לראות שהמילה susceptible נוספה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

@Reuters Children are almost immune to the coronavirus.

That's a fact.

על פי השיטה Thesaurus, חזרו 3 מתוך 4 מילים מ-keywords. התווספו vulnerable , susceptible. ניתן לראות שהמילה susceptible נוספה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט תוצאות המדדים של כל השאילתות:

search engine 4					
recall	precision @50	precision @10	precision @5	precision	Query num
0.979058	0.82	0.6	0.8	0.700375	1
0.995169	0.82	1	1	0.768657	2
0.948718	1	1	1	0.884462	3
0.987952	0.96	0.9	0.8	0.623574	4
0.652632	0.34	0.2	0.2	0.392405	5
0.983607	0.44	0.5	0.4	0.43956	6
0.393939	0.94	1	1	0.938144	7
0.981481	1	1	1	0.988806	8
0.976636	0.7	0.7	0.6	0.79771	9
0.974026	0.96	1	1	0.9	10
0.97619	0.98	1	1	0.791506	11
0.76	0.18	0.5	0.8	0.085973	12
0.928571	0.64	1	1	0.193069	13
0.954167	0.88	1	1	0.916	14
0.921569	0.84	0.6	0.6	0.824561	15
0.935484	0.98	1	1	0.805556	16
0.638298	0.94	1	1	0.909091	17
0.948718	0.84	1	1	0.74	18
0.869565	0.7	0.6	0.8	0.540541	19
0.833333	0.52	0.8	0.6	0.354839	20
0.824561	0.74	0.8	1	0.502674	21
0.932584	0.4	0.5	0.6	0.324219	22
0.987179	0.84	0.7	1	0.633745	23
0.936937	0.84	0.9	0.8	0.597701	24
0.785714	0.26	0.6	0.8	0.193833	25
0.970238	0.88	0.9	1	0.687764	26
0.962766	0.98	1	1	0.819005	27
0.641791	0.8	0.7	1	0.601399	28
0.766234	0.92	0.9	1	0.598985	29
0.807692	0.88	1	1	0.815534	30
0.938144	0.82	0.9	1	0.758333	31
0.984615	0.26	0.4	0.4	0.25098	32
0.907258	0.98	1	1	0.941423	33
0.94709	0.78	1	1	0.730612	34
0.858209	0.48	0.7	0.8	0.481172	35
0.882575	0.752571	0.811429	0.857143	0.643777	average
0.936937	0.84	0.9	1	0.700375	median
0.393939	0.18	0.2	0.2	0.085973	min
0.995169	1	1	1	0.988806	max
0.755937747					MAP

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

Search_engine_5 – Word Net

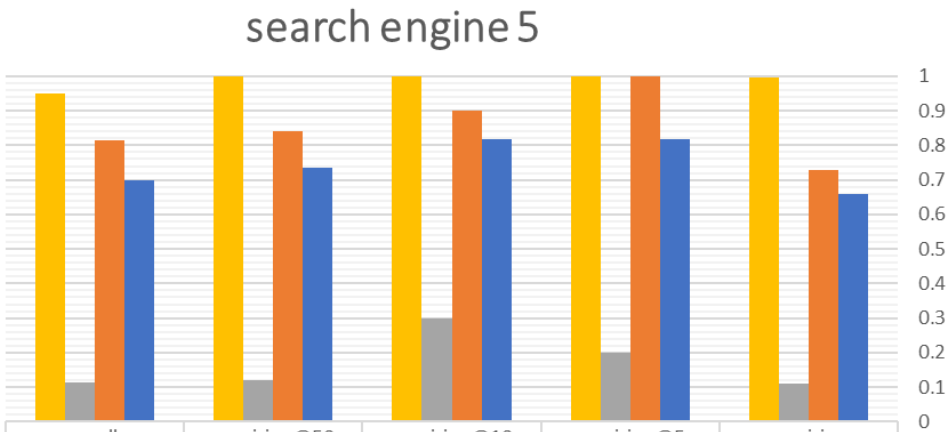
השתמשנו בשיטה WordNet בתצורת Query Expansion. השיטה משתמשת בקוד פתוח ושמו wordnet, אשר שייך ל-NLTK. השיטה מחזירה רשימה עם מילים נרדפות לכל מילה מהשאילתה.

לשם מיטוב איכות תוצאות האחזור, החלטנו לבצע Query Expansion אך ורק עם מילים העומדות בתנאים הבאים:

- המילה הנרדפת נמצאת במילון הקובץ ההופכי
- המילה הנרדפת לא נמצאת בשאילתה
- לאחר בדיקה, תבוצע הוספה ראשונית לרשימה, שממנה נדרג ונסנן את ה- **2 הנרדפות עם הציון הכי גבוה** כך:
  - **מונה השאילתה לא יעבור את 4 :** מונה סופר כמות הוספות עבור כלל השאילתה
  - **מונה המילה מהשאילתה לא יעבור את 1 :** המונה סופר את כמות המילים שהוספנו באיטרציה על מילה ספציפית, ומתאפס עבור כל מילה מחדש.

נקודה למחשבה - לשם שיפור איכות התוצאות, שילבנו את השיטה עם שיטת Global Method בעזרת החלפת דירוג המילים הנרדפות - בדירוג ה-Sij לשם הוספת המילים הללו. אך השיטה לא הניבה שיפורים משמעותיים.

פירוט המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.69947112	0.735680272	0.817142857	0.817142857	0.659403289
median	0.813432836	0.84	0.9	1	0.728813559
min	0.112903226	0.12	0.3	0.2	0.109677419
max	0.951690821	1	1	1	0.995515695

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט השאילתות:

על פי השיטה **Word Net** ביצענו **query expansion** – המילים שמסומנות באדום הן המילים שהתווספו לשאילתה לאחר ההרחבה. מכיוון שכאן הדירוג מבוצע בעזרת **cos similarity** בעזרת **tf-idf** ישנה חשיבות לכמות המילים המשתופות עם השאילתה.

שאילתה מספר 1 -

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

fauci paper hydroxychloroquine sars (keywords)

הציוצים שהוחזרו:

.1

So, Fauci reported that HYDROXYCHLOROQUINE IS EFFECTIVE AS A CURE AS WELL AS A VACCINE IN 2005!!!! ...AND IT WAS EFFECTIVE AGAINST SARS-CORONA!!! ...but he will not report this again, now, with this new pandemic? He is DEEP STATE.

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו theme,report, ניתן לראות שהמילה report נוספה.

.2

Please Share everywhere. Dr Fauci has known since 2005 that Chloroquine, Hydroxychloroquine, had killing effects on Sars-Covid diseases. Fauci penned a Report in 2005 about the Positive effects on Covid.

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו theme,report, ניתן לראות שהמילה report נוספה.

.3

Dr Fauci wrote medical paper on

Hydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he down plays it now

על פי השיטה WordNet, חזרו 4 מתוך 4 מילים מ-keywords. התווספו theme,report ,

.4

ScottAdamsSays Also, why is HCQ a "Chloroquine is a potent inhibitor of SARS @ coronavirus infection and spread," but not SARS-CoV-2? Paper is from 2005

https://t.co/QqSWDo7Erq

על פי השיטה WordNet, חזרו 2 מתוך 4 מילים מ-keywords. התווספו theme,report, ניתן לראות שהמילה report נוספה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

Fauci knew this in 2005 when he wrote a paper saying Chloroquine (even milder than Hydroxychloroquine) is practically a vaccine against ALL Coronaviruses & SARS.

Relatives of those who died from CORONA should be able to SUE HIM! He killed them!
https://t.co/WanX8YP9Vq

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו theme,report, ניתן לראות שהמילה report נוספה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

.1

Why if the death rate from Covid-19 is less than Influenza, do we need a mandatory vaccine if we don't require one for the flu?

על פי השיטה WordNet, חזרו 2 מתוך 4 מילים מ-keywords. התווספו grippe,killing,kill,influenza. ניתן לראות שהמילה influenza נוספה.

.2

@educated_educ8r 1. Kids can spread influenza pre-symptomatically just like any other virus. (!) Influenza is actually more dangerous for them; less for us.

2.Vaccines for flu ONLY work 40% of the time--more facts. Yet we are still in school.

על פי השיטה WordNet, חזרו 1 מתוך 4 מילים מ-keywords. התווספו grippe,killing,kill,influenza. ניתן לראות שהמילה influenza נוספה.

.3

ImHereForTrump @WAVY_News it may be seasonal like influenza. 20x more @ people have died from covid this year than died from influenza in 2017. If this is a thing that'll happen every year, it'll be a long ride

על פי השיטה WordNet, חזרו 1 מתוך 4 מילים מ-keywords. התווספו grippe,killing,kill,influenza. ניתן לראות שהמילה influenza נוספה.

.4

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is…

על פי השיטה WordNet, חזרו 2 מתוך 4 מילים מ-keywords. התווספו grippe,killing,kill,influenza.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

RT @Peoples_Pundit: The seasonal flu kills far more kids each year than coronavirus will ever.

It's not even close. Further, pneumonia is…

על פי השיטה WordNet, חזרו 2 מתוך 4 מילים מ-keywords. התווספו grippe,killing,kill,influenza.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

.1

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו plant,gate.

.2

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו plant,gate.

.3

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו plant,gate.

.4

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו plant,gate.

.5

RT @SBSNews: Bill Gates has confirmed he will not use a COVID-19 vaccine to implant microchips in people. https://t.co/CR22a9bkOz

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו plant,gate.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

.1

@Bogs4NY NY has reached herd immunity because they were hit so hard from the get go. Life goes on...

. **על פי השיטה WordNet, חזרו 3 מתוך 3 מילים מ-keywords. התווספו gain, resistance, exemption, crowd.**

.2

@soledadobrien NY did not “bring” their numbers down. NYC was hit so hard early on that they likely have reached herd immunity.

. **על פי השיטה WordNet, חזרו 3 מתוך 3 מילים מ-keywords. התווספו gain, resistance, exemption, crowd.**

.3

@RooneyB21 @murray_nyc @sarahcpr You must not realize that even if herd immunity were possible with this virus (which it seems like it's not), that millions of people would die before we reached herd immunity.

. **על פי השיטה WordNet, חזרו 3 מתוך 3 מילים מ-keywords. התווספו gain, resistance, exemption, crowd.**

.4

RT @richardzussman: On herd immunity, Dr. Henry says no one in the world has reached herd immunity. The goal is a vaccine. #bcpoli #covid19…

. **על פי השיטה WordNet, חזרו 3 מתוך 3 מילים מ-keywords. התווספו gain, resistance, exemption, crowd.**

.5

@NateSilver538 @TheStalwart Herd immunity was reached everywhere the virus burned out. Herd immunity is *way* below the commonly stated 60-70%.

. **על פי השיטה WordNet, חזרו 3 מתוך 3 מילים מ-keywords. התווספו gain, resistance, exemption, crowd.**

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

['child', 'kid', 'resistant']

הציוצים שהוחזרו:

.1

So Trump says children are "almost immune from this disease" and that gets taken down for claiming someone is immune from the coronavirus, which it does not say.

He clearly says they are LESS susceptible.

Why pretend you don't get that?

Why should I listen to your analysis?

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו child, kid, resistant.

.2

Facebook removes video of Trump falsely claiming children are "almost immune" to the coronavirus https://t.co/NMJfmtjQiA

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו child, kid, resistant.

.3

Facebook removes Trump's video in which he said children are "almost immune" to COVID-19. https://t.co/sBkOwndwzA

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו child, kid, resistant.

.4

Facebook removes Trump post falsely claiming children are 'almost immune' to coronavirus https://t.co/L9CjSzbeMO

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו child, kid, resistant.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

.5

Kids are not immune from #COVID19

Trump:

“If you look at children, children are almost, and I would almost say definitely, but almost immune from this disease. So few and it’s they’ve got stronger-hard to believe,I don’t know how you feel about it”

WTF?

https://t.co/cYr8PEhW3s

על פי השיטה WordNet, חזרו 3 מתוך 4 מילים מ-keywords. התווספו child, kid, resistant.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט תוצאות המדדים של כל השאילתות:

search engine 5					
recall	precision @50	precision @10	precision @5	precision	Query num
0.78534	0.84	1	1	0.731707	1
0.951691	0.84	0.9	1	0.778656	2
0.807692	0.98	0.9	0.8	0.908654	3
0.939759	0.96	0.9	0.8	0.624	4
0.8	0.38	0.3	0.4	0.41989	5
0.918033	0.38	0.4	0.2	0.430769	6
0.190476	0.916667	1	1	0.916667	7
0.822222	1	1	1	0.995516	8
0.864486	0.72	0.5	0.4	0.814978	9
0.393939	0.98	0.9	1	0.919192	10
0.933333	0.98	1	1	0.837607	11
0.2	0.357143	0.5	0.6	0.357143	12
0.404762	0.2	0.8	0.8	0.109677	13
0.9375	0.88	0.9	1	0.914634	14
0.901961	0.9	0.8	0.8	0.844037	15
0.887097	0.96	1	1	0.829146	16
0.347518	0.9	0.9	1	0.890909	17
0.882051	0.82	0.9	1	0.738197	18
0.623188	0.74	0.9	1	0.551282	19
0.409091	0.44	0.8	0.8	0.325301	20
0.54386	0.78	0.7	0.6	0.568807	21
0.730337	0.12	0.3	0.2	0.288889	22
0.948718	0.84	0.8	0.8	0.643478	23
0.891892	0.9	1	1	0.603659	24
0.410714	0.4	0.7	0.6	0.343284	25
0.934524	0.9	0.9	0.8	0.69163	26
0.909574	0.96	1	1	0.9	27
0.41791	0.78	1	1	0.615385	28
0.558442	0.86	0.9	0.8	0.581081	29
0.509615	0.86	1	1	0.828125	30
0.927835	0.76	0.9	1	0.756303	31
0.861538	0.16	0.5	0.4	0.248889	32
0.112903	0.875	1	1	0.875	33
0.910053	0.9	0.9	1	0.728814	34
0.813433	0.48	0.7	0.8	0.467811	35
0.699471	0.73568	0.817143	0.817143	0.659403	average
0.813433	0.84	0.9	1	0.728814	median
0.112903	0.12	0.3	0.2	0.109677	min
0.951691	1	1	1	0.995516	max
0.778037432					MAP

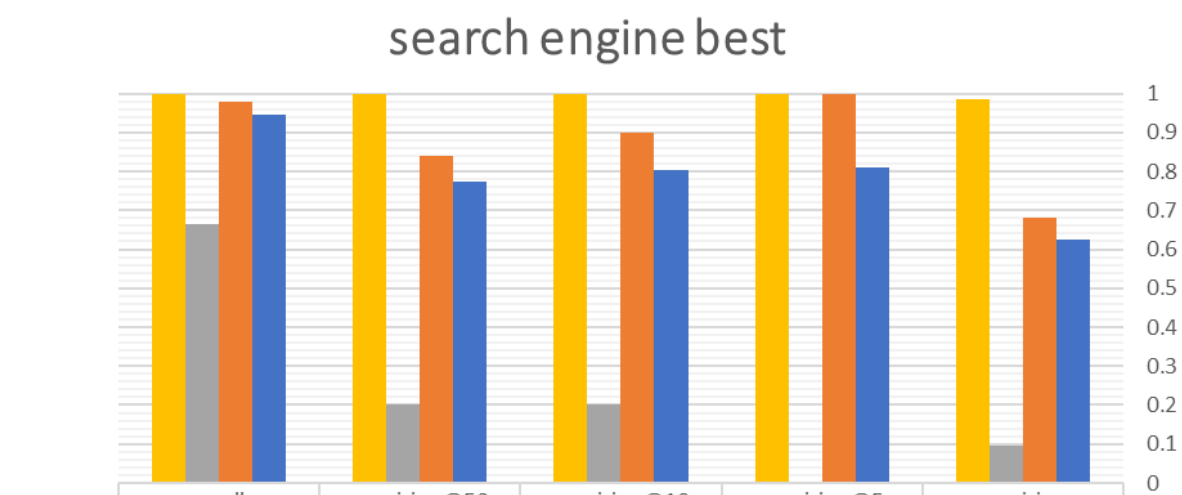
ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

Search_engine_best – Word2Vec + Spelling Correction

הפונקציות שמימשנו עבור כל אחת מהשיטות מפורטות בגישות השונות.

אופן מימוש השיטה הוא דירוג בעזרת **Word2vec** ותיקון השאילתה בעזרת השיטה **Spelling Correction**, במידה ותהיה מילה עם שגיאת כתיב בשאילתה נוכל לתקן אותה ונחליף אותה במילה המתאימה. זאת על מנת לקבל את המסמכים הרלוונטיים ביותר, ללא התיקון אנו עלולים להתעלם ממילים שכן חשובות לנו לחישוב הדירוג וקבלת המסמכים.

פירוט המדדים:



	recall	precision@50	precision@10	precision@5	precision
average	0.947103463	0.772571429	0.802857143	0.811428571	0.623498048
median	0.979166667	0.84	0.9	1	0.681818182
min	0.663157895	0.2	0.2	0	0.097165992
max	1	1	1	1	0.985347985

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

פירוט השאילתות:

על פי השיטה **Word2Vec** נשים לב שככל שהמסמך דורג במקום יותר גבוה זה אומר שהמילים שנמצאות בשאילתה והמילים שנמצאות במסמך חולקות הקשרים משותפים רבים יותר בקורפוס ולכן וקטור המסמך יהיה קרוב לוקטור השאילתה ולכן ממוקם בדירוג גבוה מבין המסמכים.

* Spelling Correction לא תרם במקרה הזה כי לא היו שגיאות כתיב בשאילתות*

שאילתה מספר 1 -

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

fauci paper hydroxychloroquine sars (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבשלושת המסמכים הראשונים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים חולקות הקשרים משותפים רבים עם המילים בשאילתה. בנוסף לכך, נראה שעבור המסמך הרביעי והחמישי אמנם הקשר מאוד דומה אך לא חזק כמו בשלושה הראשונים.

.6

RT @StoneSculptorJN: Fauci's own NIH published a paper showing that Chloroquine is a "potent inhibitor of SARS" (but there is little money**…**

.7

RT @USAMomUtah: @CindyProUSA Yes. Dr Fauci published a paper in 2005 praising the use of hydrochloroquine in the originals SARS COVID outbr**…**

.8

Dr Fauci wrote medical paper on

Hydroxychloroquine was only medicine to prevent and treat from SARS and Corona Virus ask him why he down plays it now

.9

@EvilDave_NXT @MichaelCoudrey U R LYING !!! @drdavidsamadi @zev_dr @raoult_didierOf DR ANTHONY FAUCI Said himself in 2005 that chloroquine, hydroxychloroquine are definitely an effective therapeutic and prophylactic against corona viruses check it out @realDonaldTrump @maddow @CNN @raoult_didierOf @CNN**!!!**

.10

RT @mitchellvii: Here's #DrFraudFauci in 2005 calling #Hydroxychloroquine a "wonder drug" in defeating SARS virus. https://t.co/XOSKpI1ZpX

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 2 -

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

flu kills more than covid (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שהמסמך הראשון הוא בעל הקשר החזק ביותר עבור המילים שמופיעות גם במסמך וגם בשאילתה, הן חולקות הכי הרבה הקשרים משותפים. נראה ששאר המסמכים גם יש הקשרים משותפים אך פחות מאשר המסמך הראשון.

.6

I guess COVID-19 kills the flu?

.7

@OHaraTony @QDROP8 Covid-19 IS the seasonal flu.

.8

Is COVID-19 deadlier than the flu? https://t.co/nbWx5HOqhP

.9

@ConservBlue2020 @RyanAFournier No, its not....the flu kills more people than Covid....get a grip.

.10

@itvnews Would this be the seasonal flu? Renamed as covid19?

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 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

gates implant microchips (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נראה שחזרו מסמכים שונים אך עם טקסט דומה (מכיוון שזה **RT**), המילים שמופיעות במסמך חולקות הקשרים משותפים עם המילים שנמצאות בשאילתה.

.6

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.7

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.8

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.9

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

.10

RT @CBSNews: .@NorahODonnell: “To be clear, do you want a vaccine so that you can implant microchips into people?”

Bill Gates: “No…I don't…

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 7 -

Herd immunity has been reached.

herd immunity reached (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבארבעת המסמכים הראשונים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים מופיעות בהקשרים משותפים רבים עם המילים בשאילתה. בנוסף לכך, נראה שעבור המסמך החמישי אמנם הקשר מאוד דומה אך לא חזק כמו באחרים.

.6

@yashar They probably reached herd immunity....

.7

Sure did, and they have reached herd immunity

.8

@clarkle119 @benshapiro Northeast has reached herd immunity.

.9

@Monideepa62 We have reached herd immunity to injustice.

.10

@BeckiMeister @philipoconnor How is this not herd immunity? https://t.co/54oIh9h2Cb

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

שאילתה מספר 8 -

Children are “almost immune from this disease.”

children immune to coronavirus (keywords)

הציוצים שהוחזרו:

על פי השיטה **Word2Vec** נשים לב שבכל המסמכים ישנם אותם מילים כמו מהשאילתה ובנוסף לכך המילים מופיעות בהקשרים משותפים רבים עם המילים בשאילתה.

.6

"Children almost immune from COVID"!

.7

No, children are NOT immune to Covid-19.

.8

@Reuters Children are almost immune to the coronavirus.

That's a fact.

.9

Children are Almost Immune from COVID-19

.10

@RaheemKassam @JackPosobiec Was a clip from Fox saying children are almost immune from #coronavirus

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

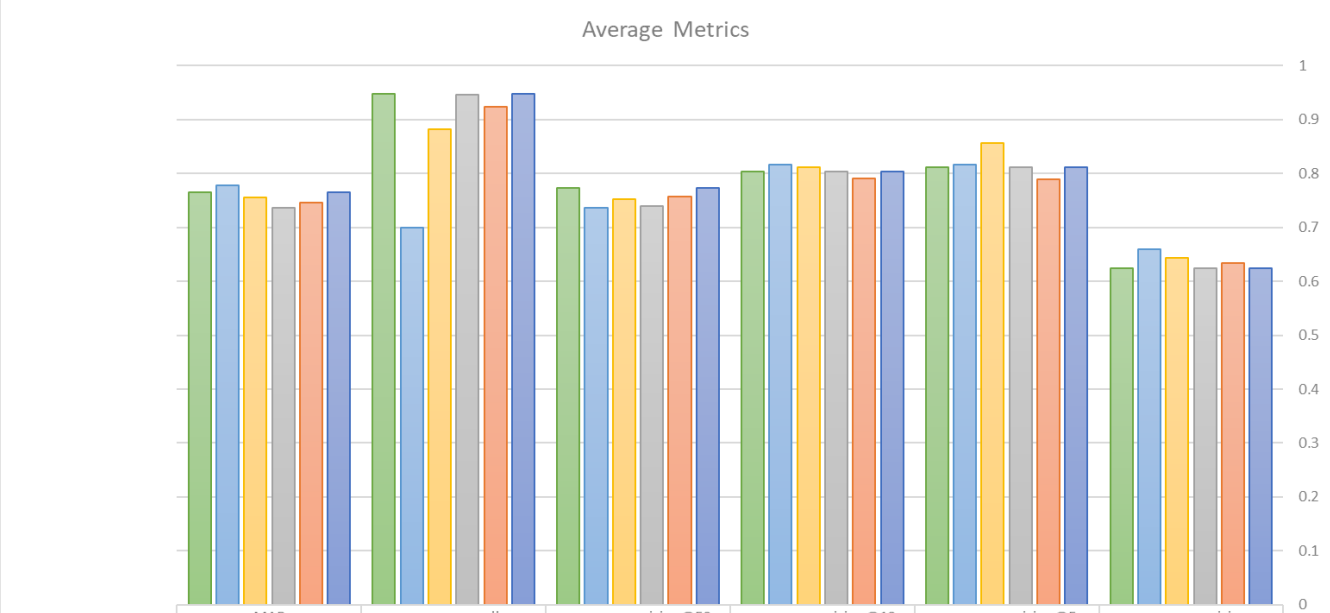
פירוט תוצאות המדדים של כל השאילתות:

search engine best					
recall	precision @50	precision @10	precision @5	precision	Query num
0.984293	0.86	1	1	0.696296	1
1	0.8	0.6	0.6	0.758242	2
0.995726	0.9	1	1	0.862963	3
0.993976	0.96	1	1	0.627376	4
0.663158	0.3	0.5	0.4	0.372781	5
0.983607	0.42	0.2	0	0.43956	6
0.831169	0.98	1	1	0.845815	7
0.996296	0.98	0.9	0.8	0.985348	8
1	0.74	0.6	0.2	0.778182	9
0.995671	0.86	0.9	0.8	0.884615	10
1	0.98	1	1	0.795455	11
0.96	0.2	0.7	1	0.097166	12
0.928571	0.54	0.9	1	0.193069	13
0.979167	0.88	0.8	0.6	0.917969	14
0.941176	0.96	0.9	1	0.806723	15
0.951613	0.86	0.7	1	0.797297	16
0.758865	0.9	0.9	0.8	0.84252	17
0.984615	0.8	0.7	1	0.747082	18
0.927536	0.8	0.7	0.6	0.528926	19
0.954545	0.8	1	1	0.276316	20
0.95614	0.84	0.9	1	0.465812	21
0.977528	0.54	0.4	0.4	0.332061	22
1	0.76	0.6	0.8	0.621514	23
0.990991	0.78	0.7	0.6	0.578947	24
0.946429	0.5	0.6	0.6	0.215447	25
0.982143	0.96	1	1	0.681818	26
0.994681	0.98	1	1	0.760163	27
0.80597	0.72	0.9	1	0.610169	28
0.811688	0.92	1	1	0.589623	29
0.951923	0.68	0.9	1	0.518325	30
0.938144	0.84	0.9	1	0.758333	31
1	0.32	0.3	0.2	0.251938	32
0.995968	1	1	1	0.94636	33
0.989418	0.94	1	1	0.730469	34
0.977612	0.74	0.9	1	0.507752	35
0.947103	0.772571	0.802857	0.811429	0.623498	average
0.979167	0.84	0.9	1	0.681818	median
0.663158	0.2	0.2	0	0.097166	min
1	1	1	1	0.985348	max
0.765074911					MAP

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

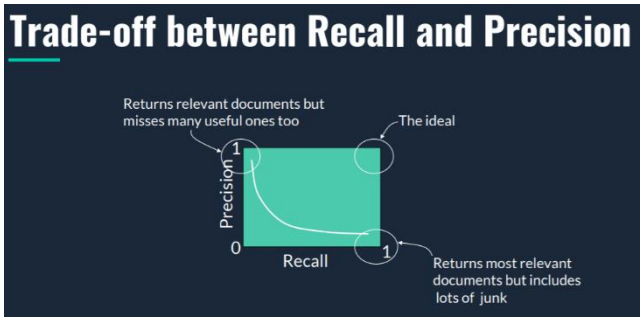
<u>בחרנו את השיטה העדיפה עבור ה – **best** מבחינת כמה מדדים:</u>

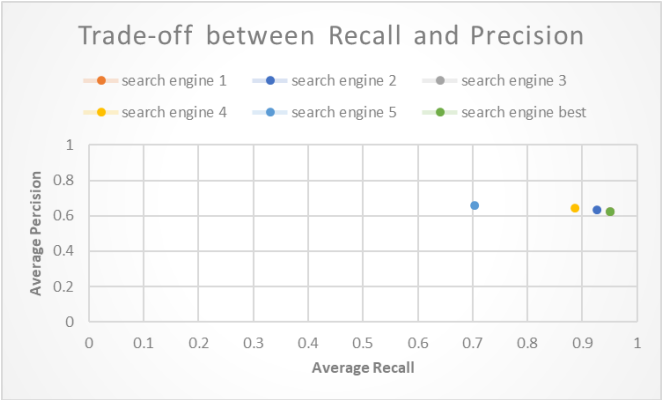
- הסתכלנו על כל מדדי הערכה שקיבלנו - MAP, average precision, average precision@5, average precision@10, average precision@50, average recall



	MAP	average recall	average precision@50	average precision@10	average precision@5	average precision
search engine 1	0.765074911	0.947103463	0.772571429	0.802857143	0.811428571	0.623498048
search engine 2	0.745140537	0.92377202	0.756571429	0.791428571	0.788571429	0.63431201
search engine 3	0.735965266	0.946663903	0.74	0.802857143	0.811428571	0.623414884
search engine 4	0.755937747	0.882575046	0.752571429	0.811428571	0.857142857	0.643777291
search engine 5	0.778037432	0.69947112	0.735680272	0.817142857	0.817142857	0.659403289
search engine best	0.765074911	0.947103463	0.772571429	0.802857143	0.811428571	0.623498048

- הסתכלנו על ה – **trade-off** בין ה – **recall** ל – **precision** כפי שלמדנו בהרצאה (התוצאה האידיאלית היא להגיע לערך 1) – חישבנו את המרחק האוקלידי בין 2 הערכים לבין **(1,1)**.





- הסתכלנו על ה – **trade-off** בין ה – **recall** ל – **MAP**– חישבנו את המרחק האוקלידי בין 2 הערכים לבין **(1,1)**.
- עשינו ממוצע בין 2 המרחקים שקיבלנו עבור כל **search engine**.

בכוונתנו למצוא את מנוע החיפוש המיטבי, שממנו נוציא את יכולות האחזור הגבוהות ביותר.
את הניתוח שלנו התחלנו בהשוואת מרחקים אוקלידים של ה – **recall** וה – **precision** מהמקסימום. לא מצאנו את סיפוקינו מהפערים בין מנועי החיפוש ולכן החלטנו להכניס מדד נוסף שיעזור לנו לבחון את הסיטואציה מחדש. את המשך הניתוח שלנו ביצענו ע"י השוואת מרחקים

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

אוקלידים של ה – **recall** וה – **MAP** מהמקסימום. מכיוון ואנחנו יודעים שמדד ה – **MAP** נותן עדיפות מסויימת ל – **precision**, נשתמש במדד משוקלל בין שני המדדים האוקלידיים הנ"ל.

average both distances	MAP, recall - distance	recall, precision - distance	Methods
0.310503141	0.240806647	0.380199636	search engine 1
0.319781769	0.266015133	0.373548405	search engine 2
0.324855655	0.269367927	0.380343383	search engine 3
0.322959465	0.270841288	0.375077643	search engine 4
0.413919732	0.373610746	0.454228717	search engine 5
0.310503141	0.240806647	0.380199636	**search engine best**

על בסיס הממוצע בין המדדים הסקנו שקבלת ההחלטות בנושא בחירת מנוע החיפוש המיטבי תוכרע ע"י הממוצע הנמוך ביותר - ראינו שהשיטה **Word2Vec** בולטת על פני כולן ולכן היא נבחרה כשיטה העדיפה ל – **best**. את ה - **Spelling Correction** בחרנו להוסיף על מנת לדייק את השאילתה.

- מתוך הבנה שהשיטות הכי טובות שלנו הינן **Global Method** ו – **Word2Vec**, שילבנו את שתי השיטות בשני חלקים שונים במנוע החיפוש: בשלב מציאת המסמכים הרלוונטיים השתמשנו בשיטת ה - **Global Method** ולמטרת הדירוג השתמשנו בשיטת **Word2Vec**. אך ראינו שהתוצאות אינן משפרות את המדדים אליהם התייחסנו. ולכן ויתרנו על שילוב זה.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

<u>פרסר</u>

הפרסר שלנו נשאר עם אותן יכולות כמו הפרסר שלנו מחלק א', רק שהפעם כן פרסרנו גם את הטוויטים שמתחילים ב-RT מכיוון שלאחר שחקרנו את הטוויטים המוחזרים לכל שאילתה, מסקנתנו היא שהרבה טוויטים מתצורה זו חוזרים כרלוונטים.

מבחינת בניית המודל של Word2Vec, שינינו מעט את הפרסר. הסרנו את היישויות, התיוגים (@). בעבור ההאשטגים, השארנו רק את המילה המפורקת (כלומר הסרנו רק את ההאשטג בתצורה המלאה שלו). ראינו שלאחר השינוי אנחנו מגיעים עם המודל לתוצאות טובות ומדויקות יותר. מכיוון שהמודל עובד לפי קשרים משותפים של מילים שהופיעו באותו חלון, השארת המילים שציינו לעיל הפריעה לדיוק של החישוב והאימון.

מבחינת המדדים השינוי המשמעותי היה בערך ה – MAP. לפני ההסרה הגענו לערך של 0.752 ולאחר ההסרה הגענו לערך של 0.765.

<u>Stemming</u>

החלטנו לפסוח על ה – Stemming מכיוון שראינו שאין שינוי משמעותי בתוצאות של המדדים.

השינוי המשמעותי היה בערך ה – MAP כאשר עם Stemming קיבלנו ערך של 0.738 וללא Stemming הגענו לערך של 0.765.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

מילון

מספר המילים במילון: 11414

המידע המוכל בכל כניסה באינדקסים:

Inverted index

מוחזק במבנה נתונים מסוג מילון כאשר המפתח הוא ה – **term**, והערך הוא רשימה כאשר הערכים בו:

Df – מספר המסמכים שהטוויט מופיעה בהם.

מילון של טוויטים – כאשר המפתח הוא מספר הטוויט והערך הוא רשימה עם הערכים:

Max_tf – מספר של המילה שמופיעה הכי הרבה פעמים במסמך.

Tf – מספר החזרות שהמילה מופיעה במסמך.

numOfUniqueWords – מספר המילים הייחודיות במסמך.

|D| - אורך המסמך.

NumOfCurrInCorpus – מספר ההופעות של מילה בכל הקורפוס.

Cij – מקדם דימיון הלא מנורמל של מילה עם עצמה במטריצת האסוציאציות.

Inverted_docs

מוחזק במבנה נתונים מסוג מילון כאשר המפתח הוא ה – **tweetID**, והערך הוא מילון של כל המילים בתוך המסמך:

המפתח הוא ה – **term** והערך הוא הנתונים שמופיעים תחתיו ב – **Inverted index**.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

<u>מפני שרצינו להגיע לזמן ריצה מיטבי בזמן החזרת התשובה לשאילתה:</u>

יצרנו את ה – Inverted docs - כאשר נקבל מה – searcher רשימה של מסמכים רלוונטים. במקום לחפש בכל ה – inverted index את הנתונים על המסמכים, נוכל לשלוף בקלות את הנתונים לכל מסמך כאשר נשמור אותם במילון תחת המפתח של מספר הטוויט.

שמירת הערכים עבור הדירוג בשלב האינדוקס – בזמן אינדוקס המילים שמרנו את הערכים שישמשו אותנו למטרת חישוב ה – tf-idf . כאשר ניגש למילה במילון כל הערכים הרלוונטים למטרת הדירוג כבר יהיו כערך במילון – מה שנותר הוא רק לבצע את החישוב.

<u>מפני שרצינו לשפר את איכות התוצאות, נעשו כמה החלטות:</u>

החזרת מסמכים רלוונטים – למטרת העלת ה – precision וניסיון להחזיר כמה שפחות מסמכים שאינם רלוונטים, מחלקת ה – Searcher תחזיר את המסמכים שאורכם גדול או שווה מאורך השאילתה (בנוסף לתנאי שבמסמך צריכה להופיע לפחות מילה אחת מהשאילתה). ראינו שתנאי זה אכן מעלה לנו את מדד ה – precision ורק במעט הוריד את מדד ה – recall ולכן ה- trade-off היה משתלם במקרה הנ"ל.

<u>כמה נקודות למחשבה שעלו לנו בעת בניית המנוע, היינו מעוניינים לנצל את הפוטנציאל כמה שיותר, אך לא ראינו שיפור משמעותי בעזרתן:</u>

שימוש בדירוג BM25 - רצינו לראות אם נדרג את המסמכים בעזרת נוסחה זו איכות התוצאות תשתפר, אך ראינו שקורה בדיוק ההפך. המדדים רק ירדו לעומת הדירוג בעזרת cos similarity.

הסרת המסמכים שמדורגים במקומות האחרונים – רצינו לנסות אופציה זו על מנת להחזיר כמה שפחות מסמכים שאינם רלוונטים במטרה להעלות את ערך ה – precision, ואכן ערך ה – precision עלה במעט, אך גרם לכך שערך ה – recall ירד משמעותית. רצינו לשמור על ערכים מאוזנים ולכן בחרנו לא להשתמש ברעיון זה.

שימוש במחלקת Searcher אשר תעבוד בצורה נוקשה יותר – התנאי הנוכחי שלנו לשליחת המסמכים הרלוונטים למחלקת ה-Ranker הינו שבכל מסמך שנשלח יהיה לפחות מילה אחת מהשאילתה. החלטנו לחקור את הכיוון של הגדלת ה-precision על חשבון ה-recall – בכך שהתנאי שבחנו הינו שמסמך רלוונטי שיישלח למחלקת ה-Ranker הינו מסמך שבו מופיעים **לפחות שתי מילים מהשאילתה.** אכן השתפר ה-precision, אך על בסיס המדדים שהגדרנו לעצמו למדידת איכות האחזור המשוקללת התנאי נפסל, לכן המחלקה הנוכחית שלנו איננה זו.

ליעד נחשון 205890320
שחר שצ'רנסקי 313326985

<u>יתרונות מנוע החיפוש שלנו על פני מנועים אחרים</u> :

<u>יתרונות</u> :

- שימוש במודל Word2Vec למטרת דירוג המסמכים – מכיוון שהמודל בנוי על בסיס הקורפוס המלא מחלק א' ישנו יותר סיכוי שנחזיר מסמכים רלוונטים יותר מאשר מנועים אחרים, מכיוון שישנו מצב שבשאילתה תהיה מילה שלא מופיעה בקובץ ההופכי, בעזרת השימוש בשיטה Word2Vec נוכל להחליף מילה שלא נמצאת בשאילת במילה הכי קרובה לה מהמודל ולא לפסוח על המילה - ובכך לדייק את האיחזור שלנו.
- בחרנו לבצע את מקסימום השיטות שניתן לבצע וכך יכולנו לבצע המון שילובים ולבחון את השיטות הנוספות – דבר שנתן לנו ידע רב בהכרת השיטות וניסיונות יצירתיים של שילובים בין כלל השיטות.
- כל המחלקות שלנו Parser, Indexer, Searcher, Ranker ניתנות לשימוש ע"י כל השיטות. כאשר מתוך ה – Search engine best נאתחל בעזרת פונקציית set כאשר נרצה להפעיל שיטה מסוימת – נאתחל ב - True. לפי האתחולים נדע לקרוא לפונקציות הרלוונטיות של אותה שיטה שרצינו להפעיל. דבר שהקל מאוד על הפעלת השיטות ועזר לנוחות הבדיקה.

<u>חסרונות</u> :

- חקירת המימוש והביצוע של Word2Vec דרש מאיתנו מאמצים גדולים של קריאה, למידה ובהמשך הרבה ניסוי וטעייה בצורת יצירת מודלים מאומנים שונים. בנוסף לכך מכיוון שהמודל היה בנוי על כל הקורפוס מחלק א' לקח המון זמן לניקוי ואימון המודל. החיסרון מתבטא בכך שיכול להיות שיכולנו לשפר את ביצועינו באחזור יותר על ידי השקעת הזמן הזה בחלקים אחרים במנוע החיפוש או בשיטות האחרות.
- שמירת נתונים כפולים גם ב – inverted index וגם ב – inverted docs, דבר שגרם לנו להחזיק יותר מידע בזכרון – לדעתנו זהו ויתור שהיה משתלם על מנת לייעל את מהירות איחזור המסמכים.
- בהינתן שימוש בשיטת Word2Vec- מעבר לכך שהמודל יחסית כבד, יש תלות ניכרת באחסון המודל בשירותי ענן - בשונה ממנועים שהשתמשו בשיטות שאינן דורשות מודל מאומן.