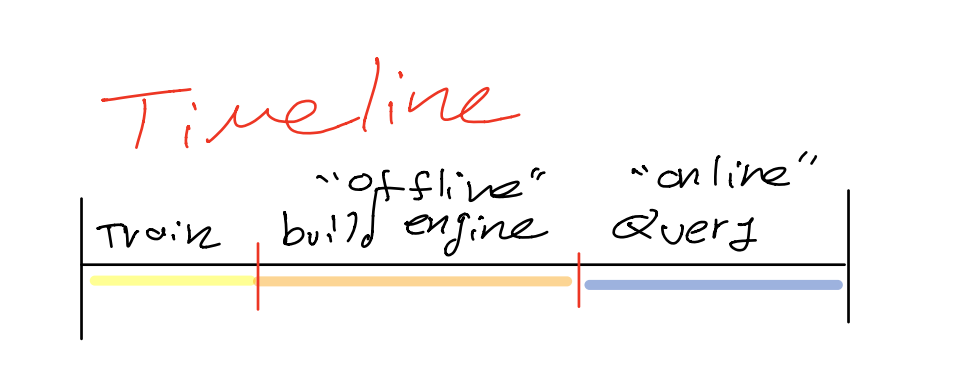
**דו" באחזור -חלק ג**

1. א. גישה מספר 2:



המנוע שלנו מתבסס בשיטה זאת במודל Word2Vec, שאנו אימנו על הקורפוס המלא (מסעיף א').

המודל מתבסס על חבילת Gensim, ובחרנו לייצר ווקטורים באורך 35, עם מספר מגבלות נוספות(לדוגמא הגבלת גודל המילון שהמודל מייצר).

יצרנו קובץ טקסט של כל הקורפוס המלא, כך שכל ציוץ עובר parsing. לאחר מכן השתמשנו בחבילה Gensim כדי לאמן את המודל על הקורפוס הקובץ הנ"ל. פעולה זאת מתבצעת באופן חד פעמי ולאחר שנוצר המודל המאומן ניתן להשתמש בו ללא צורך להריץ שוב פעולה זו שוב.

**בזמן בניית המנוע(Offline)**

אנחנו יוצרים את הinverted\_index הקלאסי שבנינו עד כה,

לאחר שהוא נבנה , אנו יוצרים שני מילונים:

א.מילון של מילים - וקטורים dic\_words– נגזר מתוך המודל המאומן, כך שרק מילים שנמצאות בinverted index וגם במודל המאומן יהיו במילון הזה ובצורה הזאת אנחנו חוסכים מקום רב בזיכרון וכמו כן בזמן הריצה בזמן הONLINE.

ב. מילון של מסמכים – וקטורים dic\_docs– לאחר שבנינו את המילון מהסעיף הקודם, נייצג כל מסמך באמצעות ממוצע הווקטורים של המילים המכילות אותו , דבר זה נעשה במהירות , בזכות העובדה שבנינו את המילון מסעיף בצורה מצומצמת.

בזמן שאילתה(Online):

א.בזמן ששאילתה מגיעה – היא עוברת parsing באותה שיטה ששאר המסמכים בקורפוס עוברים.

לאחר מכן מתבצע query expansion באמצעות מודל Word2Vec והמילון מסעיף א.

הפעולה מרחיבה את השאילתה באמצעות מציאת המילה עם הIDF הנמוך ביותר(הכי פחות נפוצה) בשאילתה.

לאחר שהרחבנו את השאילתה, נמצא את כל המסמכים שמכילים לכל הפחות מילה מתוך השאילתה המורחבת, באמצעות PostingFile, נקרא להם רלוונטיים.

חישוב הדמיון:

ניצור לשאילתה ווקטור שמייצג אותה כמו שעשינו לכל מסמך במילון dic\_docs

עבור כל מסמך 'רלוונטי' בדקנו את המרחק האוקלידי שלו מן הווקטור של השאילתה.

החישוב הזה מהווה את החלק הארי לדמיון בין המסמכים לשאילתה, ובנוסף נתנו עדיפות למסמך ככל שהוא מכיל יותר מילים השייכות לשאילתה, באמצעות ביטוי זה נתנו עדיפות למסמכים שמכילים את המילים המדוייקות מתוך השאילתה , מלבד להיותם דומים במרחב האוקלידי(סמנטית).

משיקולים שנרחיב בהמשך בחרנו שלא להחזיר את כלל המסמכים, אלא את 1000 המסמכים הראשונים.

יש לציין שמכיוון וככל שהמרחק גדול יותר במרחק האוקלידי ככה הקשר בין המסמך לשאילתה נמוך יותר, ולכן במקרה זה ככה שערך הדמיון נמוך יותר כך הקשר חזק יותר. גם את דירוג המסמכים ביצענו בלוגיקה הזאת.

ב. גישה 1: Search\_engine1

המנוע הנ"ל מתבסס על מודל GloVe.

בדומה למודל Word2Vec גם הוא משמש לייצוג מילים במרחב הווקטורי.

את מודל זה לא אימנו והשתמשנו בקובץ שקיים במערכת על ידי Stanford.

מתוך (<https://nlp.stanford.edu/projects/glove>)

שלב בניית המנוע (Offline):

תחילה נבנה הinverted\_index אשר זהה לכלל המנועים.

לאחר מכן נבנה מודל הGloVe :

הקובץ נקרא מתוך הנתיב '../../../../glove.twitter.27B.25d.txt'

בדומה למנוע מגישה 2, יצרנו שני מילונים dic\_words dic\_docs,. כל ווקטור המייצג מילה\מסמך הוא באורך 25. גם כאן שני המילונים נבנים בצורה מצומצמת על בסיס המילים שקיימות בinverted index הנוכחי. יש לשים לב כי: לדוגמא עבור מסמך שכל מילותיו לא קיימות במודל Glove(דבר זה ייתכן מכיוון והמודל לא אומן על קורפוס זה או על קורפוס דומה) אז לא יהיה ניתן לתת ייצוג לאותו מסמך , במקרה כזה הוא כלל לא ייכנס למילון.

כאשר מגיעה שאילתה חדשה, באמצעות posting file נעבור על כלל המסמכים הרלוונטיים( כלל המסמכים שמכילים לכל הפחות מילה מן השאילתה) .

עבור כל מסמך נבצע פונקציית דמיון Cosine Similarity המשתמשת בשיטת TF-IDF.

בשלב הבא עבור כל אחד מהם נבצע את המרחק האוקלידי בין הווקטור של השאילתה לווקטור של המסמך.

מלבד המרחק האוקלידי, לכל זוג (מסמך,שאילתה) ניתן משקל המבטא את חוזק הקשר למודל Glove.

ובכך התגברנו על מקרה כי מסמך שמתאים מאוד לשאילתה מבחינה סמנטית אך מרבית מילותיו אינן נמצאות במודל GloVe יאוחזר במקום טוב , זאת מכיוון והמשקל שלו עם Glove יקבל ציון נמוך והוא יפגע פחות. דוגמא נוספת היא שחלק גדול ממילות השאילתה לא נמצאות במודל ולכן השימוש במודל 'פחות משיג את המטרה'.

כמובן שהמשקל הזה מנורמל בפונקציית הדמיון הסופית.

בשלב הבא נבצע את פונקציית הדמיון המשותפת שכוללת : CosSim קלאסי (TF-IDF),מרחק ווקטורי כתלות בעוצמת הקשר למודל Glove.

ככל שהמרחק קטן יותר כך הקשר הדוק יותר, ואילו בCosSim ככל שהערך קטן יותר הקשר פחות הדוק. ולכן, השתמשנו בפונקציה שמנרמלת את המרחק ונותנת ערך גדול יותר ככל שהמרחק קטן יותר.

כמו כן נשים לב שטווח הערכים כאן הוא בין 0 ל 1 , בדיוק כמו ,Cosine similarity ובכך השגנו איזון בשילוב בין שתי השיטות.

לבסוף נדרג ונחזיר את המסמכים ככל שערך פונקציית הדמיון גדולה יותר.

**ג.** גישה 3: wordNet – באמצעות שיטה זו ניתן למצוא מילים דומות ונרדפות למילים ב-query ובכך להרחיב את ה-query הקיים ולספק תוצאות מתאימות לשאילתא גם אם לא בהכרח מכילות מילים ספציפיות מהשאילתא.

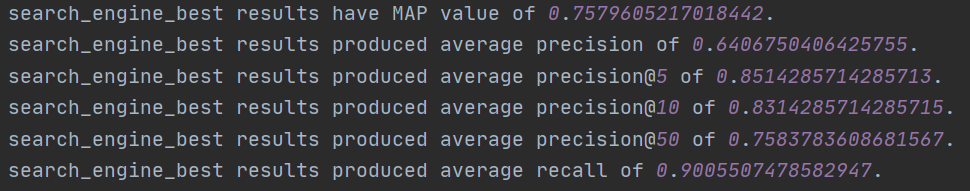
למשל עבור חיפוש המילה **flu** (שפעת) יתקבלו תוצאות המכילות את המילה **influenza.**

התוצאה השלישית שנקבל עבור חיפוש המילה הוא הציוץ:

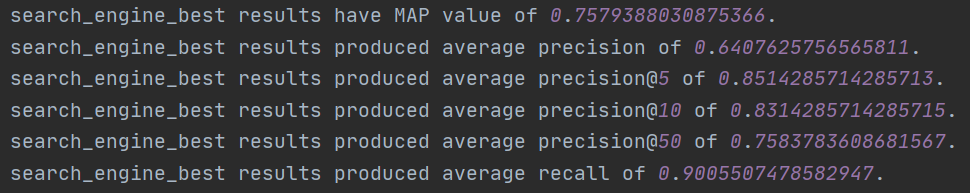
***Kids can spread influenza pre-symptomatically just like any other virus. (!) Influenza is actually more dangerous for them; less for us.***

גישה זו מחפשת מילים נרדפות עבור 2 המילים הראשונות ב-query (מתוך הנחה שאלו המילים המשמעותיות בשאילתא) ומוסיפה אותן ל-query, תוך שימוש בספריית **nltk** באלגוריתם **wordnet**.

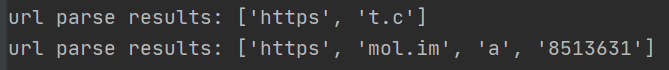
1. בחלק זה ויתרנו על הכנסת מילים מפורסרות מה-url למילון שלנו, כיוון שאינו נתן לנו כל ערך מוסף, לראיה:

תוצאות המדדים עם מילות url במילון:

ללא מילות url:



ניתן לראות כי אין כל שינוי משמעותי במדדים, אולם בכל זאת בחרנו לוותר עליו מכיוון שהוא מוסיף ל- inverted index מילים חסרות משמעות וחסרות כל ערך מוסף, כמו למשל:

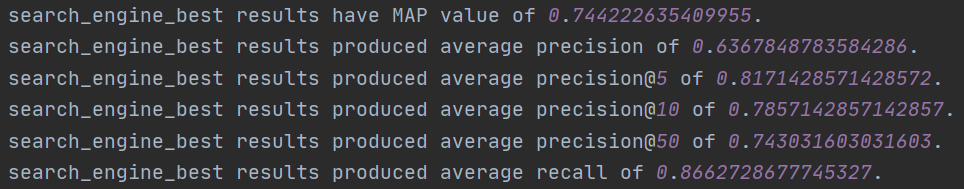


מילים אלה סביר שאף אדם לא יחפש, כלומר אלו לרוב אינן מילות חיפוש סבירות, ולכן העדפנו לוותר על הכנסה של מילים המתקבלות מפרסור של url.

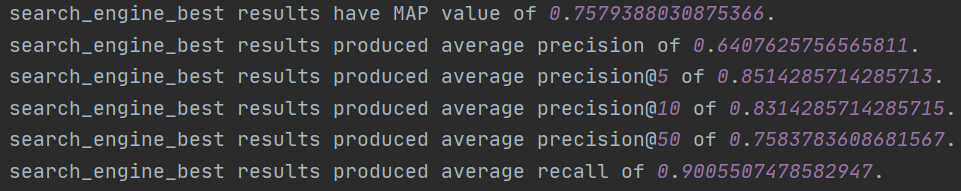
בנוסף לכך, חשוב לזכור כי כעת עבדנו על קורפוס יחסית קטן, ולכן ההשפעה של פרסור מילים חסרות משמעות והכנסתן למילון היא קטנה, אך בקורפוסים גדולים זה עלול לפוצץ את המילון שלנו במילים חסרות משמעות שאינן רלוונטיות ולתפוס נפח נוסף מהזיכרון.

Stemming:

תוצאות המדדים עם strmming:



תוצאות המדדים ללא stemming:



ניתן לראות על פי התוצאות הנ"ל הבדל יחסית משמעותי במדדים ובאיכות התוצאות ולכן בחרנו לא להשתמש בstemmer.

1. כמות המילים במילון עם מילים מפרסור url:



כמות המילים במילון בלי הכנסת מילים מפרסור url:

כמות המילים במילון עם סטמר:

ללא סטמר:



סיכום: לאחר ההחלטות להוריד מילים מפורסרות מ-url וללא שימוש ב-stemmer על פי ההסבר בסעיף הקודם, כמות המילים במילון שלנו היא 11864.

1. אנו שומרים באינדקס שלנו tupel המכיל שלושה מילונים:

( inverted index, posting, docs to info)

* Inverted index- מילון שהמפתח בו הוא term והערך כמות הטוויטים בהם הterm שהופיע בקורפוס. למשל:

במילון הזה נעשה שימוש במודלים לצורך חישוב tf-idf באמצעות חישוב הdf, משמש לsearcher.

* Posting- מילון שהמפתח בו הוא term והערך הוא רשימה של tuples כך שכל tuple מכיל 3 ערכים: (tweetId, number of occurrences of the term in the tweet, tweet words count) - כלומר מס' ציוץ, כמות המופעים של הterm בציוץ, כמות המילים בציוץ.

כמו הinverted-index גם במילון הזה נעשה שימוש ב-searchers לצורכי חישוב של cosSim.



* Docs\_to\_info – מילון המכיל מידע ברמת הtweet ולא ברמת הterm בניגוד ל-2 המילונים הקודמים. המפתח הוא מס' הציוץ והערך הוא tuple בעל 4 ערכים – מילון של כל term בtweet לכמות המופעים שלו בציוץ, כמות המילים בציוץ, תאריך הציוץ, תוכן הציוץ.

תוכן הציוץ ותאריך הציוץ אינם רלוונטיים למנוע החיפוש עצמו, ומשמשים אותנו לצרכי דיבאג – לוודא שאכן התוצאות הגיוניות (כיוון שמדובר בקורפוס קטן השארנו את התוכן הזה, כמובן שאם היה מדובר בקורפוס גדול היינו נמנעים מלשמור מידע זה על מנת לחסוך בזיכרון).

נעשה שימוש במידע במילון זה לצורך חישוב של cosSim.

the tuple for tweet id: '1280921542243659776'**({'Literally': 1, 'fuck': 1, 'many': 1, 'flu': 2, 'stats': 1, 'would': 1, 'kind': 1, 'aggressive': 1, 'testing': 1, 'every': 1, 'year': 1, 'people': 1, 'getting': 1, 'virus': 1, 'dieing': 1, 'building': 1, 'immunity': 1},** 18, 'Wed Jul 08 17:47:48 +0000 2020', 'Literally just fuck off, how many flu stats would there be if they did this kind of aggressive testing of the flu every year. If people are getting the virus and very few are dieing, we are building our own immunity. <https://t.co/0SzSQj2Dor>')

5. על מנת להביא לזמן ריצה מיטבי:

1. במודל GloVe,Word2Vec , כאשר יצרנו את המילונים dic\_words,dic\_docs, דאגנו להכניס רק מילים

אשר נמצאות בקורפוס הנוכחי (נבנה על בסיס inverted\_index).

זאת מתוך ההבנה כי מודל GloVe בכלל אומן על קורפוס אחר, ולכן לבטח יכיל מילים רבות שאינן קשורות לקונטקסט של המסמכים ושל השאילתות שיגיעו למנוע האחזור השלנו. באופן דומה Word2Vec אומן על קורפוס ענק בגודל 10Million מסמכים. ולכן גם במקרה זה מילים רבות שהופיעו בקורפוס הענקי לא יופיעו במקרה זה. זמן הריצה מיטיב פה גם עבור הOFFLINE , המילון שנבנה הוא קטן הרבה יותר, ובעיקר עבור Online ,המילון קטן הרבה יותר ולוקח פחות זמן למצוא כל ערך של מפתח(מילה). באותו אופן מסמכים שלא רלוונטיים(לא ניתנים לייצוג וקטורי מהסיבות שפירטנו מעלה) לא נכללים במילון.

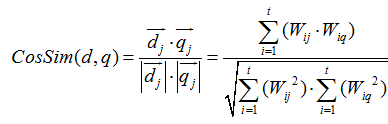
1. באימון מודל Word2Vec השתמשנו בפרמטר min\_count = 6



פרמטר זה, דואג להכניס למודל אך ורק מילים שהופיעו לכל הפחות 8 פעמים. את המספר 8 בחרנו מתוך ההבנה שגודל מילון גדול מידי יאט את המנוע שלנו , יעמיס על הזיכרון, ובעיקר יגדיל את הזמן למציאת המפתח.

1. במידה ומספר המסמכים הרלוונטיים עבור שאילתה הוא גדול מאוד(גדול מ2000 באמצעות שימוש בPosting File ), נשתמש במודל GloVe רק על k המסמכים עם ציון Cosine Similarity הגבוה ביותר.

בכך אנו חוסכים זמן רב בOnline. נזכיר שכדי להשתמש להחזיר תשובות לשאילתה אנו צריכים לבצע פעולה של חישוב מרחק וקטורי בין כל מסמך לשאילתה, פעולה זו יחסית יקרה. ולכן החלטנו לא לבצע חישוב זה עבור כל מסמך שעלול להיות רלוונטי, אלא רק עבור k המסמכים שקיבלו את הציון הגבוה ביותר עם Cosine Similarity על בסיס TF-IDF . נזכיר כי פונקציית הדמיון שלנו היא 'פיוז'ן' של שתי שיטות משקול, ולכן בין כה וכה היינו צריכים לחשב את Cosine Similarity, ולכן נשמור את התוצאה וגם נשתמש בה כמעיין 'סלקציה' ראשונית שתגיד על מי לבצע חישוב של מרחק וקטורי.

1. גודל ווקטור במודל Word2Vec הוא 35. בחנו וראינו כי כאשר גודל הווקטור גדול מ35, איכות התוצאות לא משתפרת באופן שניתן להבחין, אך זמן הריצה עולה.
2. שימוש במבני נתונים מתאימים, לדוגמא עבור איסוף מסמכים רלוונטיים השתמשנו במילון מתוך ההבנה כי בהמשך נצטרך לגשת שוב באופן נקודתי לאותם 'מפתחות'.
3. חיסכון בקריאה לפונקציות. נתבונן לדוגמא בפונקציות Cosine Similarty על בסיס TF-IDF

ניתן לחשב באופן חד פעמי לשאילתה

ידוע מראש לכל מסמך

ניתן לראות שאת המכנה לדוגמא ניתן לחשב ברגע שהגענו למסמך ואין צורך לחשב אותו מספר פעמים. בפעם הראשונה בלבד שראינו מסמך שעוד לא עברנו עליו, נחשב עבורו את המכנה(סכום המשקלים של השאילתה ידוע ללא תלות בכלל במסמך, את סכום המשקלים של המילים במסמך ניתן למשוך מתוך המילון המכיל את כלל המסמכים והמילונים שלהם.

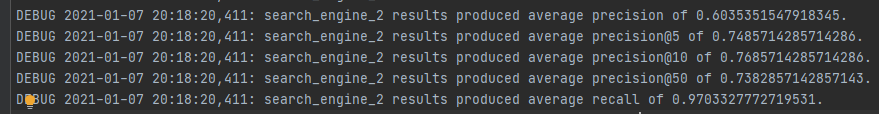
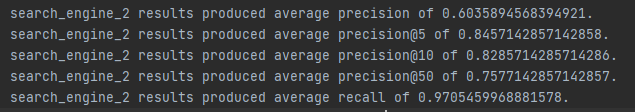
את המונה נגדיל בצורה איטרטיבית(בכל פעם שהמסמך יוזכר מתוך המעבר על Posting file ).

6. שיפור התוצאות

1. כדי לשפר את איכות התוצאות הגדלנו את גודל הוקטור של המודל המאומן Word2Vec מ25 ל35

(תוצאות אמפיריות) לפני ואחריי:

25:

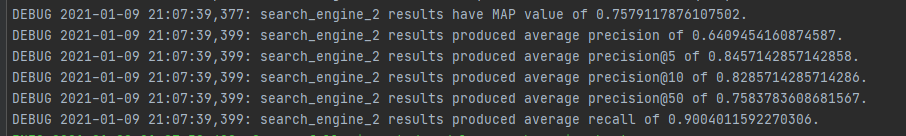


1. עבור מודל GloVe הבנו כי לא מספיק אך ורק לייצג מסמכים באמצעות הווקטורים , ולכן ביצענו פונקציית דמיון משולבת עם Cosine Similarity על בסיס tf-idf. התוצאות השתפרו אך עדיין המודל לא מאומן על הקורפוס ולכן התוצאות היו פחות טובות ובחרנו לאמן מודל Word2Vec
2. גם במודל Word2Vec הבנו כי לא מספיק להביע קשר של מסמך-שאילתה על בסיס הוקטורים בלבד ולכן השתמשנו גם בפונקציית שנותנת 'חיזוקים' עבור כל מסמך.

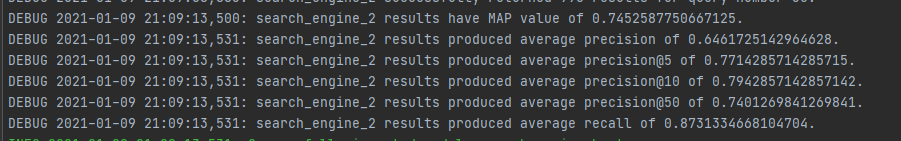
לפנינו ההבדלים בתוצאות:

def relevant\_and\_cosSim\_vecs\_with\_bonuses(self, query\_as\_list):

כאשר נותנים בונוס של 0.12 למילה משותפת.



כאשר משתמשים בפונקציית מרחק 'טהורה':

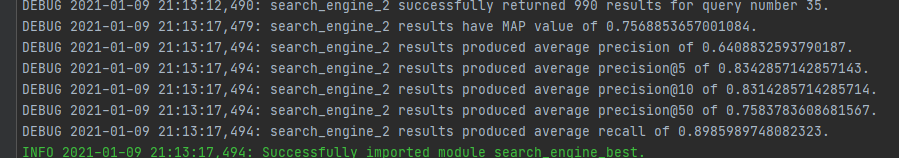


יש לציין שניסינו גם להעזר בפונקציות דמיון אחרות כגון Cosine Similarity ו bm25 , אך רק הפונקציה הנאיבית הזאת היא שהביאה לנו את התוצאות הטובות ביותר, ולכן החלטנו להשתמש בה.

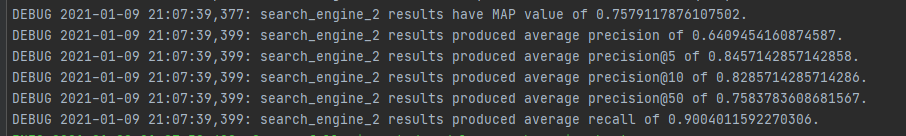
1. בשיטה Word2Vec מלבד השימוש במרחק הווקטורי , נעזרנו במודל גם כדי להגדיל את השאילתה.

לפנינו ההבדלים בין שימוש במודל עם הרחבת השאילתה לבין שימוש ללא הרחבה.

ללא הרחבה:



עם הרחבה:

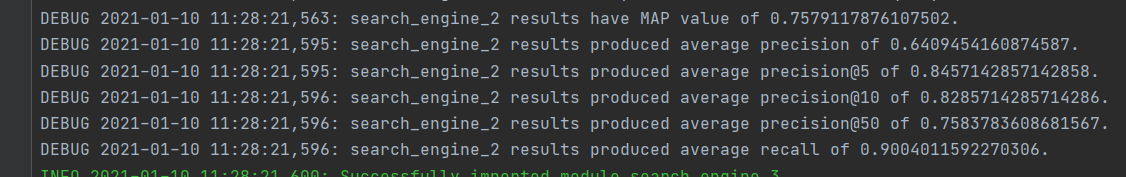


1. אנו בחרנו שלא להשתמש בStemmer, מכיוון שהוא גורע מהתוצאות משמעותית

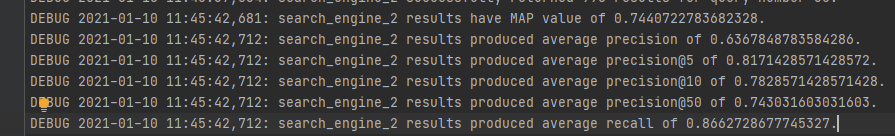
ריצה על מנועים 1,2,3 ללא סטמר.

DEBUG 2021-01-10 11:28:34,605: Total runtime was: 0:00:44.440319

תוצאות של 2 (best ) ללא סטמר:



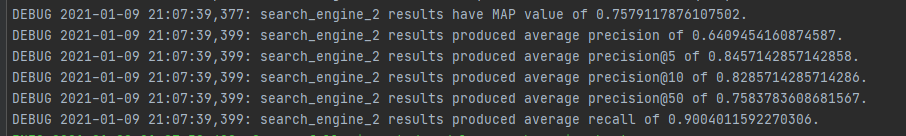
תוצאות של 2 (best ) עם סטמר:



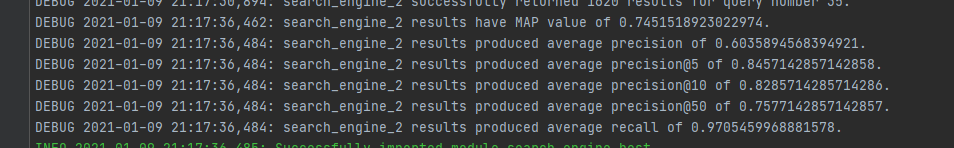
1. הוחלט להגביל את החזרת התוצאות למקסימום 990 התוצאות הטובות ביותר.

ניסינו לדמיין את המודל כמנוע חיפוש אמיתי שמשתמשים ברחבי העולם ישתמשו בו, בעודנו משתמשים בתדירות גבוה מאוד במנוע החיפוש כמו גוגל, לא יצא לנו מעולם להגיע מעבר לעמוד 5-6 מכל סיבה שהיא. ניתן לראות גם את השיפור במדדים:

עם הגבלה ל990 מסמכים:



ללא הגבלה:

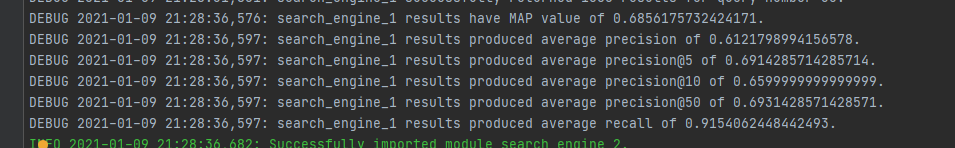


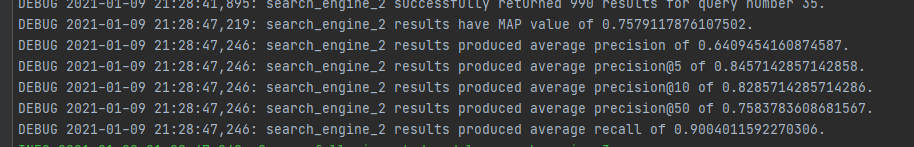
ובכן – נראה שמדד הrecall אכן נפגע , אך שוב , אנו מתייחסים למסמכים שחזרו מתחת למקומות ה1000 כמסמכים לא רלוונטיים כלל שהמשתמשים לא יגיעו אליהם , ולכן ה"אובדן" במדד הrecall לא מייצג אובדן אמיתי. מלבד זאת , כלל המדדים השתפרו.

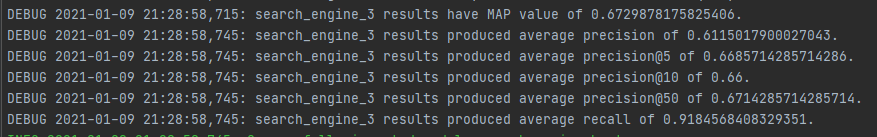
7. החלטת השיטה המיטבית.

על החלטת השיטה המיטבית ביצענו בצורה אמפירית לחלוטין. לאחר נסיונות רבים ואין סוף וריאציות ניסינו להביא לרמה הכי גבוהה כל אחד מן המנועים, כל זאת תוך כדי שמירה על זמני ביצוע ריאליים.

אלו התוצאות שקיבלנו :







בחרנו שbest יהיה מנוע מספר 2 המתבסס על Word2Vec.

כלל המדדים חשובים ומצביעים על תכונות טובות\פחות טובות שיש למנוע, אך בחרנו לתת את מירב ההתייחסות למדד של precision @5 . מתוך האמונה כי המסמכים הראשונים הם שמעידים על טיב המנוע, ושוב, מסמך שחוזר במקום ה-200 במנוע אחד וחוזר במקום נמוך יותר או לא חוזר כלל במנוע אחר, לא באמת יכול להעיד לטובה מכיוון והרוב המוחלט של המשתמשים לא יגיעו לתוצאות האלו לעולם ומבחינתנו הם שקופים.

לסיכום:

מדד MAP : מנוע 2 עדיף באחוזים בודדים

מדד Recall : אין שוני מובהק בין המנועים

מדד AVG Percision : מנוע 2 עדיף באחוזים בודדים.

**מדד Percision at 5 : מנוע 2 עדיף באופן מובהק בעשרות אחוזים (66%,84%,69%)**

8. שימוש בAPI חיצוני :

א. שימוש בgensim.models השתמשנו בחבילה Word2Vec באמצעותה אימנו את המודל, וטענו אותו לזיכרון.

ב. שימוש בNumPy על מנת ליצור ווקטורים למסמכים,מילים.

ג. שימוש בPandas על מנת לאפשר קריאה של קבצי פרקט המייצגים את הקורפוס.

ד. שימוש בScipy בחבילת spatial שמסייעת בחישוב מרחק אוקלידי בין שני וקטורים.

ה.שימוש בnltk בחבילת WordNet , שמסייעת בהרחבת השאילתה, ובמציאת המילה הכי דומה למילה נתונה. כמו כן בחבילת TweetTokenizer , PorterStemmer שמסייעות בParsing על הקורפוס והשאילתה.

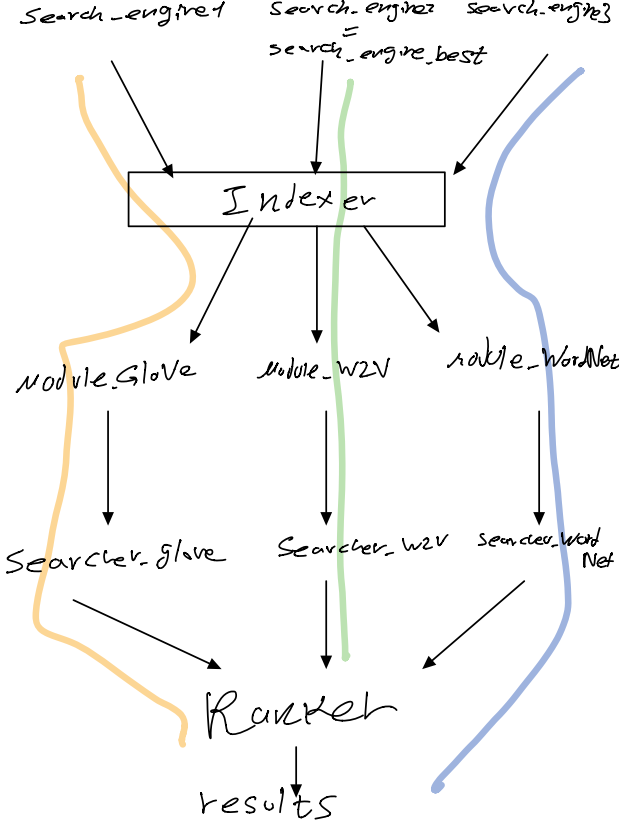
ו.שימוש בemoji לפרסור Emoji’s.

החבילות הללו עזרו לנו מאוד בכל מה שקשור במציאת דמיון, חישוב מרחקים וקטוריים , פעולות פשוטות של אלגברה לינארית , שיפור יכולת הParsing ועד אימון מודלים שלמים בשורת קוד.

עם הקלות והנוחות שבשימוש חבילות אלו, אנו בעצם יוצרים תלות בחבילות הללו ובמידה וישתנו הן עלולות לפגוע ביכולות האחזור שלנו. לדוגמא שימוש Emoji – החבילה הזאת מייצגת אימוג'י באמצעות טקסט אינפורמטיבי.במידה ויחליטו להכניס לחבילה ערכים חדשים , או לדוגמא לתאר אימוגי יחיד על ידי מספר רב של מילים – דבר שכזה יכול לפגוע משמעותית ביכולת האחזור. לדוגמא אם יחליטו שהאימוגי ☹ - יהיה במקום sad -> “depressed sad face cry” המסמך יכול לקבל משמעות אחרת לגמרי ממה שהתכוון הכותב.

ולכן על מנת לשמור על המנוע מעודכן, ועם יכולת אחזור משתפרות, יש לשים לב לעדכונים השותפים של הapi השונים.

9.הFlow של התוכנית שלנו.



חשוב לנו לציין שביצענו המון וריאציות שונות עד שהגענו למנועים שביצענו, וכדי להראות את התהליך הותרנו לדוגמא פונקציות שמימשנו ולא השתמשנו בהם בקוד:

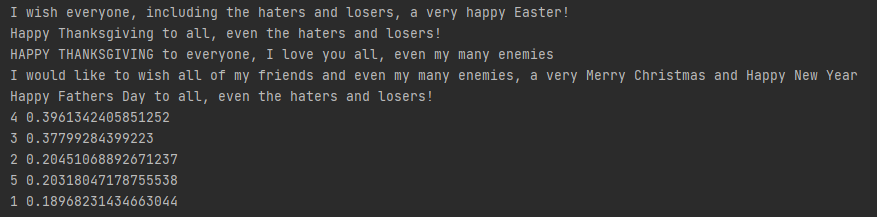
לדוגמא BM-25 , סוגי פונקציות דמיון שונות. בנוסף המצאנו מספר פונקציות דמיון "משלנו" כדי שיענו על צרכים ובעיות מסוימות, כמו לדוגמא פונקציה הופכית למרחק מכיוון שהשתמשנו בה ביחד עם פונקציית Cosine Similarity שמבטאת יחס הפוך.

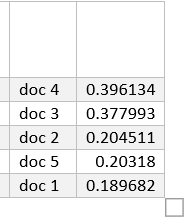
בנוסף פונקציות רבות ששימשו אותנו באופן חד פעמי לאימון מודל הWord2Vec, נותרו בקוד במידה ונרצה לאמן שוב.

כתבנו לעבודה מודול שלם של טסטים, כולל לחלקים רבים שלParser ואפילו לחישוב פונקציית דמיון.

השתמשנו גם בידע תיאורטי שיש לנו מהעבודה בחלק ב', שם נדרשנו לחשב ולדרג מסמכים על פי

TF-IDF, הבאנו את התיאוריה לפרקטיקה,ובכך ווידאנו גם שתשובתנו בחלק ב' נכונה ובעיקר שפונקציית הדמיון שכתבנו בפייתון בה אנו משתמשים רבות נכונה.

(לשם כך יצרנו את הקורפוס בגודל 5 מסמכים, בנינו את הStopwords שיהיה בהתאם לשאלה מחלק ב').



התשובות יצאו זהות לחלוטין.

10. בחינת ציוצים 1,2,4,7,8 עבור כל מנוע:

**search\_engine\_best( +2):**

תוצאות עבור שאילתה 1 – " **Dr. Anthony Fauci wrote in a 2005 paper published in Virology Journal that hydroxychloroquine was effective in treating SARS**":

1 . RT @USAMomUtah: @CindyProUSA Yes. Dr Fauci published a paper in 2005 praising the use of hydrochloroquine in the originals SARS COVID outbr…, Tweet id: 1291133082976899072

2 . @Macro\_\_Business @TheRISEofROD @IngrahamAngle @realDonaldTrump “Chloroquine is a potent inhibitor of SARS coronavirus infection and spread” - The Virology Journal, 2005 (The official publication of Dr. Fauci’s NIH)

https://t.co/nipfRlfZCv, Tweet id: 1290535411433926656

3 . @TomFitton @realDonaldTrump Is it true Fauci wrote in virology journal in 2005 that hydroxychloroquine functions as a cure and a vaccine??, Tweet id: 1291138937017307136

4 . Published online 2005 Aug 22. #Virology journal publishes study showcasing the positive effects of #chloroquine aganist #SARS type #coronavirus #FauciTheFraud knew #Hydroxychloroquine works #Scamdemic #COVID19 #depopulation #Corruption, Tweet id: 1285169749333356544

5 . 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, Tweet id: 1291683169557721088

ניתן לראות כי עבור שאילתה זו חמשת התוצאות הראשונות שחזרו הן מכילות את מילות ה-query, למשל בתוצאה הראשונה שחזרה כל המילים מופיעות ואכן הציוץ שחזר עונה במדויק על השאילתה, וכך ניתן לומר גם על הציוץ השני שחזר. בנוסף ניתן לראות עבר תוצאה 4 כי רובו מורכב מהאשטאגים של המילים במופיעות בשאילתה ועדיין חזר בין הראשונים מה שמראה כי פרסר שלה-hashtag מצליח להביא תוצאות מיטיבות. fauci מופיעה כחלק מה-hashtag וכן גם המילה Hydroxychloroquine מה שמראה כי עצם פרסור ה-hashtag עזר לנו במציאת תוצאות רלוונטיות.

תוצאות עבור שאילתה 2 – " **The seasonal flu kills more people every year in the U.S. than COVID-19 has to date**" :

1 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291791311998029824

2 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291802650980343809

3 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291786635097657344

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?, Tweet id: 1291615710330880000

5 . RT @tylerblack32: A typical seasonal flu kills about 300-600,000 people per year in the world. COVID-19 was at ~300k worldwide in May 15th,…, Tweet id: 1283768736340557824

ניתן לראות כי הציוצים שהתקבלו עונים כולן על תוכן שאילתה ורלוונטיים למענה עליה- ציוצים 1-3 הם RT וכיוון שהטקסט של הציוץ המקורי כלול בתוך הציוץ של הRT והוא מפורסר כחלק מהציוץ קיבלנו תוצאות זהות והגיוניות עבור שלושת המקומות הראשונים מה שמראה גם על עקביות בציונים והתוצאות.

עבור ציוצים 4 ו5 נראה כי הם עונים יפה על השאילתה וכן רוב מילות השאילתה מופיעות בהם ונוסף לכך גם ההקשר לשאילתה מדוייק.

תוצאות עבור 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**":

1 . Philanthropist and Microsoft founder Bill Gates has denied that his support for coronavirus vaccine research is cover for his plans to dominate the world with 5G-activated mind control microchips. In an interview with CBS News’ Norah O’Donnell, Gates w... https://t.co/vD2iy9sHXv, Tweet id: 1286582747239047168

2 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 12836622886549176323 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283579019154399232

4 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283609352755007489

5 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283586103983910912

ניתן לראות כי קיבלנו ציוצים שאכן מכילים את כל מילות השאילתה ועונות על תוכן השאילתה, כלומר מחזירות תוצאות רלוונטיות. הציוץ שחזר במקום הראשון מספק תשובה אינפורמטיבית מלאה ואיכותית עבור השאילתה. עבור ציוצים 2-5 כל הציוצים זהים, אך שונים, כלומר עצם כך שהקורפוס שלנו הוא טוויטר וכל ציוץ נחשב למסמך, אז גם RT הוא ציוץ לכל דבר ולכן אם הקורפוס שלנו מכיל הרבה ציוצים שהם RT אנחנו נתייחס אליהם כמסמך לכל דבר (כלומר לכל אחד יש ID משל עצמו) ונאחזר אותם במידה ויהיו רלוונטים. המקרה הנ"ל מראה בצורה ברורה שהציוץ עליו עשו RT הוא מאוד רלוונטי ולכן מכיוון שהטקסט עליו עשו RT הוא חלק מהציוץ נאחזר את כל הציוצים האלה שהם בעצם RT.

תוצאות עבור שאילתה 7 – " **Herd immunity has been reached**":

**1 .** Sure did, and they have reached herd immunity, Tweet id: 1287465471449559040

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

https://t.co/BznRjRcT17, Tweet id: 1285619546573942785

**3 .** @AndrewM94635548 @AFen93317147 @CrankyChris916 @pilot1996 That's cause they reached herd immunity., Tweet id: 1282423869438033922

**4 .** @ellymelly Have we reached herd immunity for the common cold?, Tweet id: 1282144460076273664

**5 .** @Monideepa62 We have reached herd immunity to injustice., Tweet id: 1281605225489379329

ניתן לראות כי התוצאות שחזרו מספקות מאוד שכן כל הציוצים שחזרו מכילים את כל מילות שאילתה וכולם מדברים על נושא השאילתה ועל אותו הקשר.

תוצאות עבור שאילתה 8 – "**children immune to coronavirus**":

**1 .** No, children are NOT immune to Covid-19., Tweet id: 1291243586835472384

**2 .** "Children almost immune from COVID!", Tweet id: 1291180630315868162

**3 .** Children are Almost Immune from COVID-19, Tweet id: 1291188776493080576

**4 .** Children are nearly immune from Covid disease!, Tweet id: 1291335639544336384

**5 .** @TeamTrump says children are almost immune to covid19. Unfortunately, these dead children weren't. https://t.co/IRioWn6kic, Tweet id: 1291376513917616128

גם כאן ניתן לראות כי התוצאות אכן מספקות וכולן מדברות על נושא השאילתה ומהותה. כל הציוצים מכילים את מילות השאילתה ומספקים את מילות החיפוש. ניתן לראות גם כי למרות שמילת החיפוש בשאילתה היא coronavirus התוצאות שחזרו מכילות וריאציות שונות למילה, כמו: Covid-19, COVID, COVID-19, covid19, כלומר בכל ציוץ שחזר coronavirus נכתב אחרת, אך זה לא הפריע לתוצאות וזאת בזכות פרסור הcovid שכתבנו, המתייחס לכל וריאציות בהן ניתן לכתוב covid19.

**search\_engine\_1:**

תוצאות עבור שאילתה 1 – "**fauci paper hydroxychloroquine sars**":

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, Tweet id: 1291683169557721088

2 . RT @RichHiggins\_DC: Fauci wasn’t “wrong.”

Fauci was “lying.”

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…, Tweet id: 1281212739482988544

3 . RT @RichHiggins\_DC: Fauci wasn’t “wrong.”

Fauci was “lying.”

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…, Tweet id: 1281210940554387467

4 . RT @RichHiggins\_DC: Fauci wasn’t “wrong.”

Fauci was “lying.”

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…, Tweet id: 1281222206861774849

5 . RT @RichHiggins\_DC: Fauci wasn’t “wrong.”

Fauci was “lying.”

Fauci knew in 2005 that HCQ worked on Coronaviruses.

Fauci will now be…, Tweet id: 1281214097854533638

ניתן לראות כי הציוץ הראשון שחזר מתאים לשאילתה וכן מכיל את כל מילות השאילתה ולכן גם דורג במקום הראשון. עבור ציוצים 2-5 ניתן לראות כי גן הם רלוונטים לשאילתה. ציוצים אלה דומים כיוון שהם למעשה RT וכל RT שמכיל את טקסט רלוונטי יוחזר. ניתן לומר גם כי המנוע שלנו אכן עקבי בתוצאות שכן כל הRT שהם רלוונטים והם זהים בתוכן חוזרים זה אחר זה.

תוצאות עבור שאילתה 2 – " **The seasonal flu kills more people every year in the U.S. than COVID-19 has to date** " :

1 . @MarkMeadows there is a vaccine every year for the flu there is not for Covid-19., Tweet id: 1291122148870938634

2 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291791311998029824

3 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291802650980343809

4 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291786635097657344

5 . mAnY pEopLe DiE fRoM tHe fLu eVeRy yEaR, Tweet id: 1283982102761844736

ניתן לראות כי הציוצים שחזרו מכילים את מילות השאילתה וכן עונים עליה. ניתן לראות כי תוצאות 2-4 הן תוצאות של RT עם אותו טקסט ולכן חזרו, מה שמראה על עקביות של המנוע.

תוצאות עבור 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** "

1 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283662288654917632

2 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283579019154399232

3 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283609352755007489

4 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283586103983910912

5 . Bill Gates, Tweet id: 1283869042458136578

בתוצאות שלפנינו ניתן לראות שמסמך 1 מכיל כמות מילים משותפות רבה מאוד, והסמנטיקה בהחלט משביעה רצון. כמו כן, חלק ניכר מהציון הגבוה של הציוץ הוא בזכות השמות שלנו בזיהוי ישויות, כמו Bill Gates , Microsoft.תוצאות 1-4 בעלות אותו "תסמין החזרתיות" המאפיין את הקורפוס שלנו בכך שיש מסמכים רבים עם אותו fulltext זאת עקב הפונקציונליות של Twitter לבצע RT.

הציון של מסמכים אלו נקבע בין היתר בזכות CovidParser אותו מימשנו, וגם בזכות HashtagParser שבא כאן לידי ביטוי לדוגמא ב #CoronavirusPandemic שמתפרק בין היתר לcoronavirus , pandemic ושניהן נמצאות בתוך השאילתה ולא כחלק מהHashtag.

תוצאות עבור שאילתה 7 – " **Herd immunity has been reached** ":

1 . @WillowNBirch Herd immunity, Tweet id: 1281118488996839424

2 . DO YOU HAVE HERD IMMUNITY?, Tweet id: 1281113303415164930

3 . @GregAbbott\_TX Herd immunity, Tweet id: 1282068791011291138

4 . @RepThomasMassie Herd immunity., Tweet id: 1282313081830305794

5 . Herd immunity, Tweet id: 1282518455200227328

ניתן לראות כי כל התוצאות שחוזרות מכילות את מילות השאילתה ומתכוונות לאותה משמעות, חוזרים ציוצים שונים בעלי תוכן זהה אך עם תיוגים אחרים, מה שמראה כי המנוע עקבי בתוצאות שהוא נותן.

תוצאות עבור שאילתה 8 – " **Children are “almost immune from this disease** "

1 . Children all almost immune..., Tweet id: 1291161553115455494

2 . "Children almost immune from COVID!", Tweet id: 1291180630315868162

3 . Children are Almost Immune from COVID-19, Tweet id: 1291188776493080576

4 . No, children are NOT immune to Covid-19., Tweet id: 1291243586835472384

5 . RT @nealrogers: “If you look at children, children are almost — and I would almost say definitely — but almost immune from this disease. Th…, Tweet id: 1291781195366203394

ניתן ראשית לשם לב, כי כל אחת מן התוצאות שלפנינו מכילה את Children,almost,immune. כל הציוצים שחזרו עונים בצורה כזו או אחרת על השאילתה

דבר נוסף הוא שניתן לראות כי from אשר מופיע בשאילתה לא זיכה את הציוץבמקום ה1 לעומת ציוץ 2, 3 בציון, זאת מכיוון ואף על פי ש from מופיע בהם ובשאילתה, הוא stop word ולכן לא בא לידי ביטוי בציון.

**search\_engine\_3:**

תוצאות עבור שאילתה 1 – " **Dr. Anthony Fauci wrote in a 2005 paper published in Virology Journal that hydroxychloroquine was effective in treating SARS**":

1 . @TomFitton @realDonaldTrump Is it true Fauci wrote in virology journal in 2005 that hydroxychloroquine functions as a cure and a vaccine??, Tweet id: 1291138937017307136

2 . RT @Asja\_jacobs: @Macro\_\_Business @Klok34702568 @EvanAKilgore @realDonaldTrump Dr Fauci new about this since 2005 he published the studies!…, Tweet id: 1288837174436286466

3 . @Macro\_\_Business @TheRISEofROD @IngrahamAngle @realDonaldTrump “Chloroquine is a potent inhibitor of SARS coronavirus infection and spread” - The Virology Journal, 2005 (The official publication of Dr. Fauci’s NIH)

https://t.co/nipfRlfZCv, Tweet id: 1290535411433926656

4 . 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, Tweet id: 1291683169557721088

5 . RT @TheMangoMan7: Fascist Fauci: Hydroxychloroquine Is 'Not Effective' in Treating the Coronavirus https://t.co/iclfVGQesj, Tweet id: 1288845290926215174

ניתן לראות כי כל הציוצים שחזרו אכן רלוונטי למהות ותוכן השאילתה, רוב מילות השאילתה מופיעות בציוצים למעט stopwords וכן פרסור מספרים עזר לנו למצוא ולדרג גבוה מסמכים שכוללים את שנת 2005.

תוצאות עבור שאילתה 2 – " **The seasonal flu kills more people every year in the U.S. than COVID-19 has to date** " :

1 . @MarkMeadows there is a vaccine every year for the flu there is not for Covid-19., Tweet id: 1291122148870938634

2 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291791311998029824

3 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291802650980343809

4 . RT @snopes: ❌ No, the seasonal flu does not kill more people every year in the U.S. than COVID-19 has to date. https://t.co/eiyBjfZsTM, Tweet id: 1291786635097657344

5 . mAnY pEopLe DiE fRoM tHe fLu eVeRy yEaR, Tweet id: 1283982102761844736

ניתן לראות כי התוצאות מספקות את השאילתה ומילות השאילתה מופיעות בכל הציוצים. ניתן להיבן כי הציוץ הראשון שחזר נעזר במילים מהרחבת השאילתה ולכן החזיר את התוצאה שהחזיר שלמשל לא כוללת את המילה Kill.

תוצאות עבור 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** "

1 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283662288654917632

2 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips

“Globalists" support…, Tweet id: 1283579019154399232

3 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283609352755007489

4 . RT @Aliciastarr001: @TheRealEddyTalk The #CoronavirusPandemic is a cover for a plan to implant trackable microchips.

“Globalists" support…, Tweet id: 1283586103983910912

5 . Bill Gates, Tweet id: 1283869042458136578

ניתן לראות כי תוצאות 1-4 שחזרו הן כולן תוצאה של RT עבור אותו ציוץ אשר עונה על השאילתה בצורה מיטבית. בנוסף ניתן לראות כי תוצאות פרסור הhashtag מביאות לתוצאות טובות שכן coronavirusPandemic פורק יפה לשתי המילים המרכיבות אותו שנמצאות בתוכן השאילתה. ציוץ 5 פחות אינפורמטיבי אך כן כולל את הישויות של השאילתה מה שמגיע מתוך זיהוי הישויות שלנו.

תוצאות עבור שאילתה 7 – " **Herd immunity has been reached** ":

1 . @WillowNBirch Herd immunity, Tweet id: 1281118488996839424

2 . DO YOU HAVE HERD IMMUNITY?, Tweet id: 1281113303415164930

3 . @GregAbbott\_TX Herd immunity, Tweet id: 1282068791011291138

4 . @RepThomasMassie Herd immunity., Tweet id: 1282313081830305794

5 . Herd immunity, Tweet id: 1282518455200227328

ניתן לראות כי הציוצים שחזרו מכילים את מילות השאילתה ללא מילות stopwords ולכן קיבלו ציונים גבוהים ודורגו גבוה יחסית.

תוצאות עבור שאילתה 8 – " **Children are “almost immune from this disease** "

1 . Children all almost immune..., Tweet id: 1291161553115455494, Score: 0.7065334719574662

2 . "Children almost immune from COVID!", Tweet id: 1291180630315868162, Score: 0.7022415153060371

3 . Children are Almost Immune from COVID-19, Tweet id: 1291188776493080576, Score: 0.6837397233373775

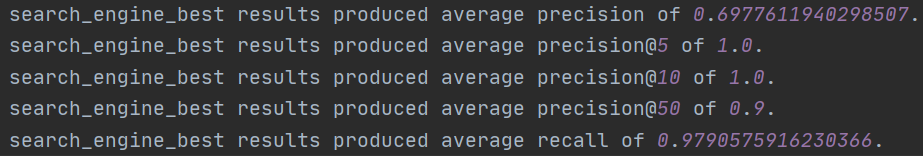
4 . No, children are NOT immune to Covid-19., Tweet id: 1291243586835472384, Score: 0.5486328610214443

5 . RT @nealrogers: “If you look at children, children are almost — and I would almost say definitely — but almost immune from this disease. Th…, Tweet id: 1291781195366203394, Score: 0.49642029297032997

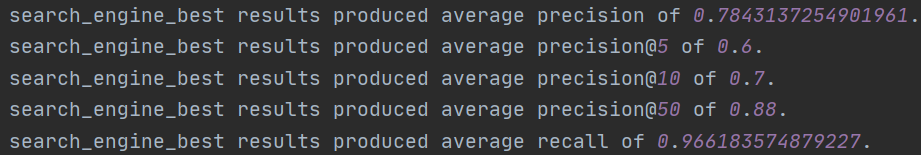
ניתן לראות כי הסיבה שציוצים אלה חזרו היא שמילות השאילתה מופיעות בציוצים וכן הקונטקסט זהה, וכן המילה disease "תורגמה" לCOVID19 על מגוון צורות הכתיבה שלו. הסיבה שציוץ 2 דורג אחרי ציוץ 1 למרות שמילה from מופיעה בו היא כיוון שהיא stopword ולכן היא לא באה לידי ביטוי בציון שניתן למסמך.

11. **מדדים עבור search\_engine\_best:**

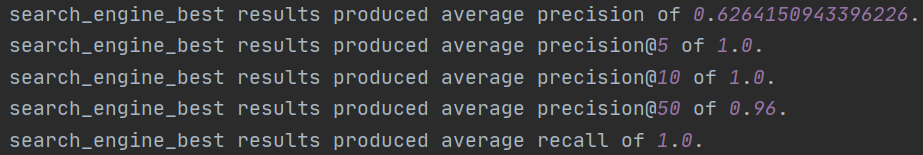
עבור שאילתה 1:



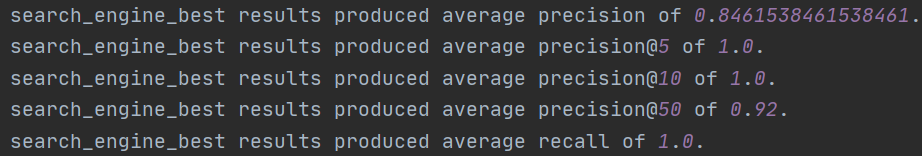
עבור שאילתה 2:



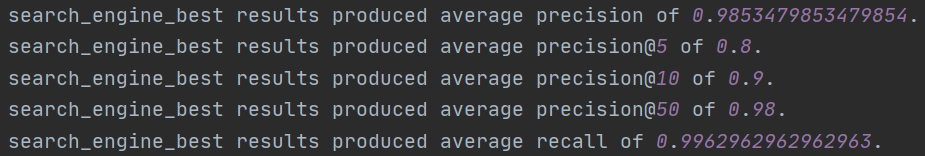
עבור שאילתה 4:



עבור שאילתה 7:



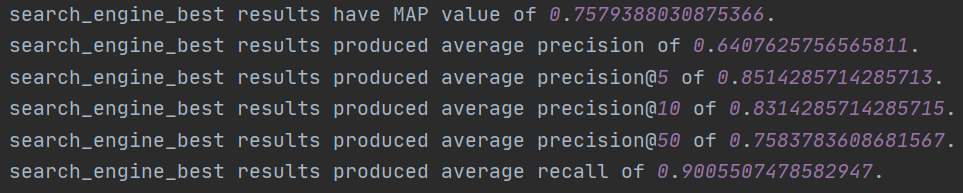
עבור שאילתה 8:



נתונים עבור שאילתות אלה:

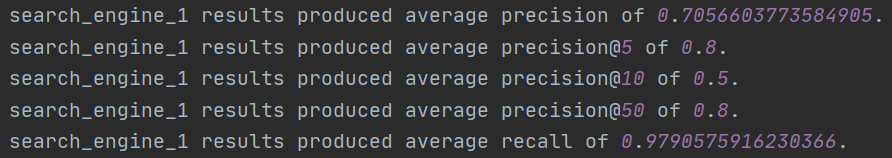
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **precision** | [**precision@5**](mailto:precision@5) | [**precision@10**](mailto:precision@10) | [**precision@50**](mailto:precision@50) | **recall** |
| **avg** | 0.787992 | 0.88 | 0.92 | 0.928 | 0.988305 |
| **mean** | 0.786151 | 0.94 | 0.96 | 0.924 | 0.992297 |
| **min** | 0.62641 | 0.6 | 0.7 | 0.88 | 0.966183 |
| **max** | 0.98534 | 1 | 1 | 0.98 | 1 |

עבור כלל השאילתות:

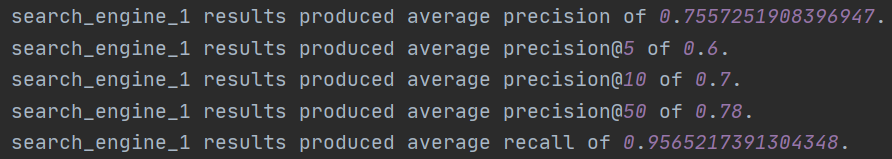


**מדדים עבור search\_engine\_1 (using Glove):**

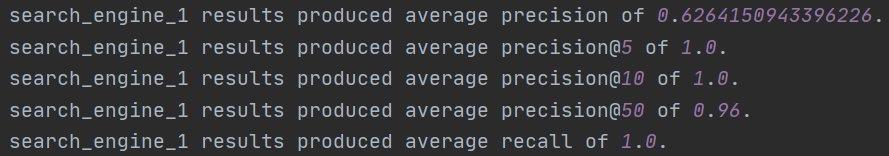
שאילתה 1:



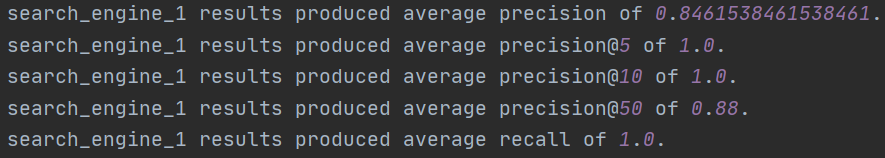
שאילתה 2:



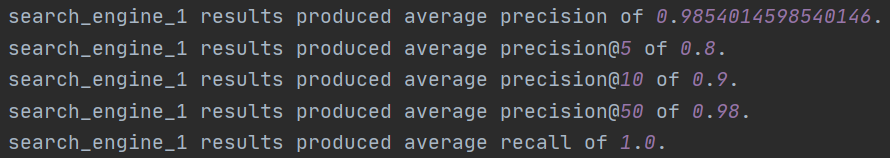
שאילתה 4:



שאילתה 7:



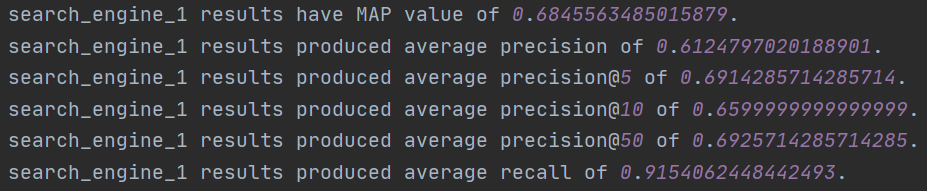
שאילתה 8:



נתונים עבור שאילתות אלה:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **precision** | [**precision@5**](mailto:precision@5) | [**precision@10**](mailto:precision@10) | [**precision@50**](mailto:precision@50) | **recall** |
| **avg** | 0.783868 | 0.84 | 0.82 | 0.88 | 0.987114 |
| **mean** | 0.75572 | 0.8 | 0.9 | 0.88 | 1 |
| **min** | 0.62641 | 0.6 | 0.5 | 0.78 | 0.95652 |
| **max** | 0.9854 | 1 | 1 | 0.98 | 1 |

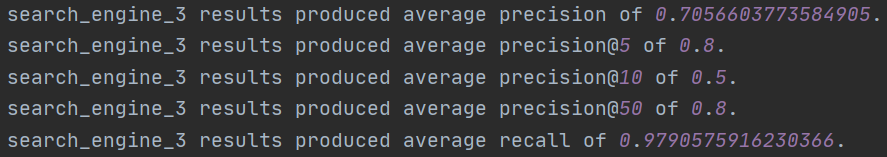
עבור כלל השאילתות:



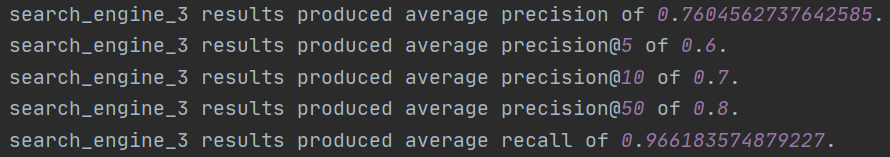
\*עבור search\_engine\_2 נקבל אותם מדדים בדיוק כמו ב-best כיוון ש- engine 2 הוא ה-best שלנו.

**מדדים עבור search\_engine\_3 (wordnet):**

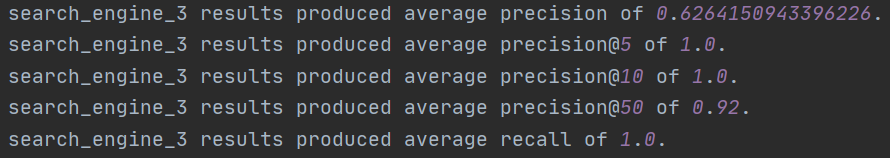
שאילתה 1:



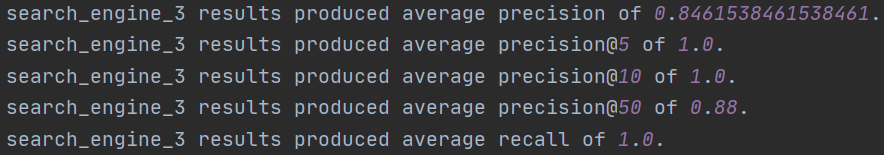
שאילתה2:



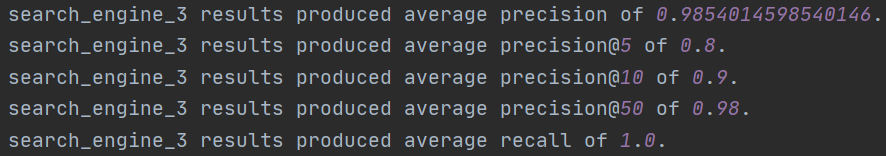
שאילתה4:



שאילתה 7:



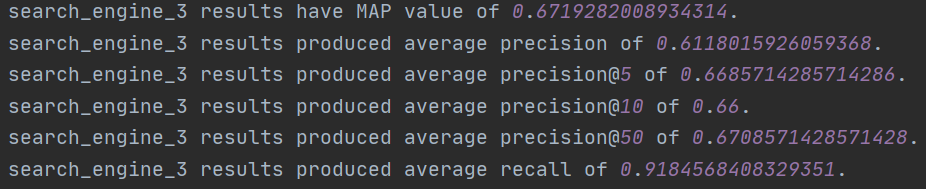
שאילתה8:



נתונים עבור שאילתות אלה:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **precision** | [**precision@5**](mailto:precision@5) | [**precision@10**](mailto:precision@10) | [**precision@50**](mailto:precision@50) | **recall** |
| **avg** | 0.784814 | 0.84 | 0.82 | 0.876 | 0.989046 |
| **mean** | 0.76045 | 0.8 | 0.9 | 0.88 | 1 |
| **min** | 0.62641 | 0.6 | 0.5 | 0.8 | 0.96618 |
| **max** | 0.9854 | 1 | 1 | 0.98 | 1 |

עבור כלל השאילתות:



סיכום מדדים של כל מנוע עבור כלל השאילתות:

12. יתרונות:

1. החוקים הנוספים שהוספנו- subdivisions&covid19 ו-emojis, מוסיפים פונקציונליות נוספת למנוע החיפוש שלנו. במעבר על הData שלנו מצאנו תצורות רבות למגפה שסובבת אותנו אך מבחינה סמנטית לכולם משמעות זהה, כמו כן תוכן רב נמצא בתוך הEmojis ,ובאמצעות השימוש בחוק תוצאות שכאלה תחזורנה.
2. מודל הword2vec שלנו אומן על הקורפוס המורחב שלנו, על מנת לספק תוצאות מיטיבות עבור השאילתות המבוססות על ציוצי קורונה ועל הקורפוס שקיבלנו, לעומת מודל שלמשל אומן מראש על ויקיפדיה ולא תואם את הקורפוס הייחודי שלנו.
3. פונקציונליות שהוספנו במנוע למציאת מסמכים המכילים אך ורק stop words תאפשר למצוא מאגר מסמכים שלם, שקיים במנוע שלנו ולא יהיה קיים במנוע שלא מימש את החוק הזה ספציפית.

חסרונות:

1. המנוע שלנו המנוע שלנו מסתמך על מרחק וקטורי ונותן בונוס למסמכים שמכילים הרבה מילים מתוך השאילתה כלומר, בנוי על שיטות מידול של bag of words, כלומר מתבסס על מילים המופיעות בquery ובקורפוס, אך אינו שם דגש על תוכן השאילתה וה-context שלה, ולכן ישנן שאילתות שהתוצאות שיוחזרו אינן רלוונטיות לשאילתה, אך הן מכילות הרבה מילים מהשאילתה ולכן יוחזרו במקומות יחסית גבוהים. אמנם המנוע שלנו מצליח להתגבר על העניין הזה הרבה בזכות מודלים מאומנים ושיטות חישוב כמו cosinSimilarity אך הוא עדיין אינו מנוע מיטבי ביחס למנועים אחרים שכן מתייחסים לתוכן השאילתה.
2. אין שימוש ב-stemmer וכתוצאה מכך הרבה מילים עלולות להתפספס אם הן אינן נכתבו בדיוק כמו בשאילתה – covid kills התוצאות שחוזרות מתייחסות רק עבור killsת למשל:

*I guess COVID-19 kills the flu?*

ולא יוחזרו תוצאות המתייחסות למילה Kill למרות שהינן רלוונטיות, כמו למשל:

*@dtb479353 @Junedocherty2 @GeeMinx @STVNews Covid doesn’t kill less than the flu. That’s just a straight up lie.*

1. במנוע שלנו אין התייחסות לתאריכים ולזמנים, כלומר אנחנו לא מפרסרים תאריכים ולא משתמשים בזה על מנת לאחזר ציוצים, מה שגורם לכך שאולי נקבל תוצאות שמכילות הרבה מילים מהשאילתה ורלוונטיות במידת מה לשאילתה, אך לא בהכרח רלוונטיות מבחינת הזמן, למשל ציוצים ישנים ולא רלוונטיים שלא נותנים כל ערך מוסף עבור השאילתה. כלומר אין שום התייחסות לכרונולוגיית הזמן מה שעלול לפגוע באיכות התוצאות.