**דו"ח- מנוע אחזור חלק ג'**

בחלק זה מימשנו שיטות נוספות לחיפוש ובנוסף שיפרנו את ביצועי המנוע. בדו"ח זה נפרט את הגישות והשיפורים הנ"ל.

1. גישות החיפוש שבחרנו לממש בחלק זה הן:

* **Spell correction –** שיטה זו מקבלת את השאילתה ובודקת עבור כל מילה האם יש בה שגיאת איות. במידה ונמצאה שגיאה היא מתקנת אותה ומחליפה את המילה השגויה במילה המתוקנת. שיטה זו אמורה להועיל לאחזור במצבים שבהם השאילתה נכתבה עם שגיאות באיות משום שהיא תדע לתקן את השגיאות ולהחזיר בכל זאת מסמכים רלוונטיים שלא היו חוזרים במידה והמילה/ המילים השגויות לא היו עוברות תיקון. לשיטה זו ישנן מספר אפשרויות לשימוש. כדי להתאים במידה הרבה ביותר למנוע שלנו ולמידע המאוחסן בו, בחרנו להשתמש בשיטה שאוספת עבור כל מילה שנחשבת כשגיאה, רשימה של מילים אפשריות לתיקון, כך שהמילה היא במרחק אות אחת מהמילה הנתונה. בדקנו את המילים שהתקבלו במילון המילים שבאינדקס שלנו ובחרנו מבין המילים את המילה שחזרה מספר מקסימלי של פעמים במילון (כלומר היא כנראה התיקון הכי רלוונטי לשאילתה).
* **Word net**- שיטה זו "מרחיבה" את השאילתה, כלומר מוסיפה לה מילים שהיא מגדירה כרלוונטיות לכל מילה בשאילתה. השיטה מתבססת על מאגר בן 144 אלף מילים באנגלית. בתור התחלה אנו עוברים על כל מילה בשאילתה, מוצאים לה את המילים המתאימות ביותר מתוך המאגר ואז מבצעים חישוב לדימיון בין המילים. לאחר מספר ניסיונות, בכדי לייעל באופן מקסימלי את ההוספה מצד אחד, ולא לגרום ליותר מידי מידע לא רלוונטי להתווסף מצד שני, החלטנו להשתמש בציון סף לדימיון שיעמוד על 0.8. בנוסף, כדי לא להעמיס במילים, אם מצאנו יותר משלוש מילים בעלות ציון הסף, נמיין את המילים שהנוספות ומתוכן נבחר את השלוש בעלות הציון הגבוה ביותר.
* **Advanced parser-** מטרת שיטה זו היא לבצע פרסור יעיל יותר של המידע, בעזרת הרחבת מחלקת Parse ע"י חוקים או עקרונות נוספים שיובילו לאחזור איכותי יותר. בשיטה זו, הוספנו מספר חוקים עבור מושגים שונים שחוזרים על עצמם בהרבה מסמכים בקורפוס והם בעלי משמעויות זהות. החוקים שבאמצעותם הרחבנו את המחלקה:

1. Mask – החלטנו לאגד את כל המילים שמתייחסות למילה "מסכה" תחת term אחד. כלומר, המילים mask, masks, n-95, kn-95 אוגדו תחת הביטוי mask וכך הגדלנו את המסמכים שיכולים להיות רלוונטיים עבור הביטוי הנ"ל.

2. United States - שמנו לב שההתייחסות לארה"ב חוזרת בהמון מסמכים במילים שונות (USA, US, United States) ולכן החלטנו לאחד את כל ה-terms בעלי משמעות זו תחת USA. למרות שהמילה us יכולה להיות בעלת משמעות אחרת לגמרי, ברוב המוחלט של המקרים הכוונה כשמתמשים בקורפוס שלנו היא לארה"ב.

3. Donald Trump- נשיא ארה"ב. במעבר על המילון שהתקבל משני הקורפוסים- הקורפוס הגדול מחלק א וגם הקורפוס הנוכחי, ובאופן כללי, ע"פ ההיגיון שלנו, הטוויטר מלא במסמכים שמדברים על ההתנהלות של נשיא ארה"ב עם הקורונה. מסמכים וביטויים רבים מאוד מכילים התייחסות לדונלד טרמפ בשמות שונים, ולכן איגדנו ביטויים כדוגמתDonald Trump, Trump, President Trump וכו' תחת הביטוי Donald Trump .

4. בנוסף לחוקים החדשים שהוספנו עבור הביטויים המוזכרים מעלה, החלטנו לבצע שיפור נוסף לחלק של פרסור Hashtags והכנסנו את הביטויים (1-3) ובנוסף את הביטויים המתייחסים לCovid-19 שביצענו בחלק א כחוקים נוספים בשלב פרסור האשטאג. כלומר, אם לאחר פירוק ההאשטאג התקבלה רשימה של מילים שמכילה את אחד הביטויים שהזכרנו, גם הם ייכנסו כאזכור ל-terms אלו.

* **Global Method**- נדרשנו לממש שיטה זו כשיטה למימוש חיפוש בחלק א של הפרויקט. שיטת חיפוש זו מייצרת מטריצה עבור כל המילים (או חלק מהמילים) בקורפוס ומחשבת את הקורלציה בין כל שתי מילים בקורפוס. בשלב החיפוש, מתבצעת הרחבת המטריצה ומתווספות לשאילתה מילים שנמצאו כבעלות קורלציה גבוהה לביטויים בשאילתה. לאחר ניסיונות רבים ומגוונים ובדיקה של אפשרויות רבות לבניית המטריצה בגדלים שונים ובצורות שונות, הגענו למסקנה כבר בחלק הקודם שבניית מטריצה מספיק גדולה שתייצג את הקרבה בין כל הביטויים בקורפוס היא בעייתית, דורשת שעות רבות (מעל שלושה ימים) של בנייה, ובמבחן התוצאה לא נראו שיפורים משמעותיים ביעילות. בחלק זה ביצענו את כל ההתאמות הנדרשות על מנת לבנות את המטריצה ע"פ ההוראות החדשות שקיבלנו בחלק זה, וגם כאן התוצאות לדעתנו לא השתפרו לעומת חיפוש ללא שיטת חיפוש מיוחדת.

1. חלק מהשיפורים ביכולות המנוע התבצעו במחלקת Parse. ביצענו מספר שינויים על מנת להקטין את מילון המילים של האינדקס כדי שיוכל להימצא במלואו בזיכרון בכל זמן הריצה. השינויים שביצענו:
2. שמות וישויות – בעבר הפרסר שמר כל רצף של שתי מילים או יותר שמתחילות באות גדולה כביטוי אחד (ע"פ הנדרש בסעיף שמות וישויות) אך שמנו לב שגם טוויטים שכתובים באותיות גדולות בלבד נשמרו כביטויים יחידים וכך התפספסו מילים שעשויות להיות משמעותיות לאחזור. בחלק זה עבדנו על זיהוי ופרסור מילים שכתובות באותיות גדולות והן לא חלק משם או ישות, וויתרנו על הביטוי המלא כי הגענו למסקנה שהרוב המוחלט של הביטויים לא מועילים לאחזור איכותי ולרוב אין בהם שימוש כלל. כך גרמנו להקטנה יחסית של המילון וגם להגדלת הרלוונטיות של מילים מסוימות.
3. URL- בחלק הקודם ביצענו parse על כל חלק בURL שמופרד ע"י '\' והסרנו סימנים רבים שלא היו רלוונטיים בעינינו. נוצר מצב שכאשר הסתכלנו בתוצאות נשמרו ביטויים רבים (מאוד) שהיו בליל אותיות ומספרים ללא שום משמעות. הגענו למסקנה שביטויים אלו כנראה לא יידרשו להימצא בשום שאילתה ולכן התעלמנו מהם בשלב הparse.
4. עבור הקורפוס הגדול (שקיבלנו בחלק א') המילון שלנו מכיל 786,016 מילים

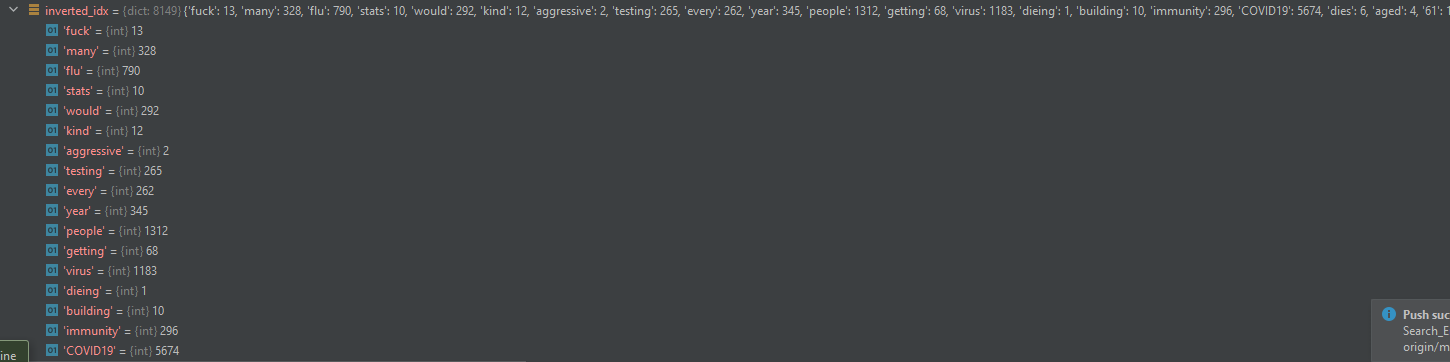
עבור הקורפוס הקטן (benchmark dataset) המילון מכיל 8149 מילים

1. האינדקס שלנו בנוי משלושה מילונים:

* המילון הראשון הוא מילון הביטויים inverted\_idx. המילון בנוי כך:

Key- הביטוי

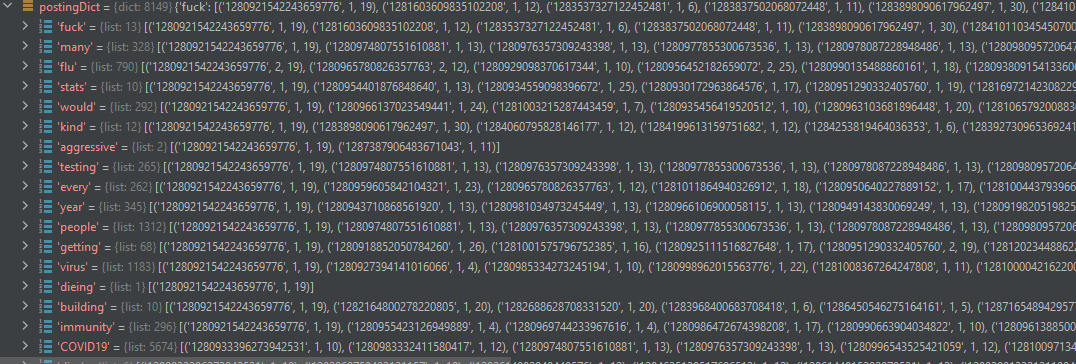
Value- מספר ההופעות של הביטוי בכל הקורפוס



* המילון השני הוא מילון הפוסטינג – postingDict שמכיל עבור כל הביטויים במאגר את המסמכים שבהם הביטוי מופיע, את מספר המופעים של הביטוי במסמך ואת אורך המסמך. המילון בנוי בצורה הבאה:

Key- הביטוי

Value- טאפל שמכיל: מזהה המסמך, מספר המופעים של הביטוי במסמך, אורך המסמך.



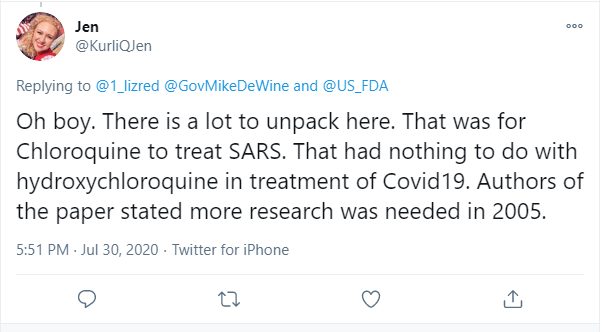
* המילון השלישי הוא מילון המסמכים- מילון זה מכיל את כל המסמכים במאגר. עבור כל מסמך נשמרים: max-tf (מספר ההופעות של המילה הנפוצה ביותר במסמך), התאריך שבו פורסם המסמך, כמות המילים הייחודיות (שמופיעות פעם אחת) המילון מיוצג בצורה הבאה:
* Key- מספר המסמך
* Value- טאפל שמכיל: max-tf, Date, num of unique terms



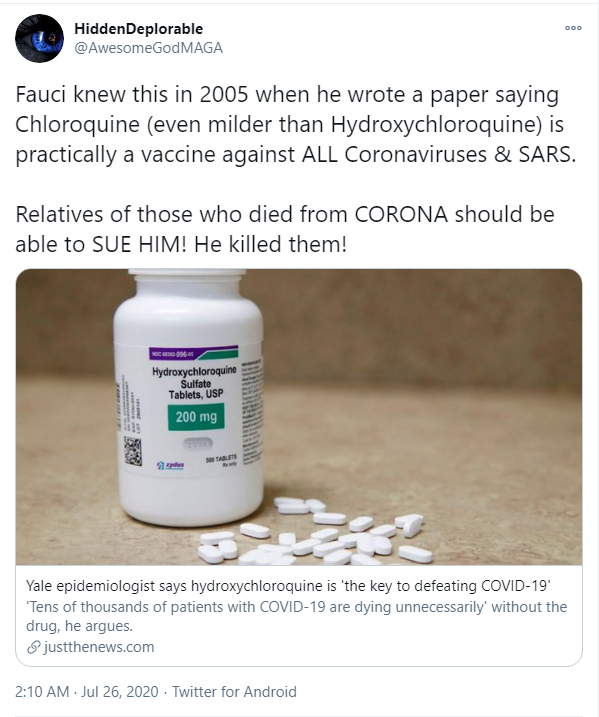
1. כפי שהסברנו בסעיפים קודמים, במהלך ריצת המנוע היו נכנסים לאינדקס המון ערכים חסרי משמעות לחלוטין, לאחר צמצום התוצאות הגענו למצב שמתקבלות תוצאות איכותיות ומועילות וכמובן שזמן הריצה הצטמצם משמעותית. בנוסף, העובדה שכל האינדקס נשמר בזיכרון במהלך כל ריצת התוכנית מייעל מאוד את הזמנים, אין צורך בקריאות מרובות מהדיסק ובחיפוש ארוך של ערכים, כל המידע זמין בזיכרון וזמן הריצה התקצר משמעותית.
2. על מנת לשפר את איכות התוצאות ניסינו לנרמל ערכים שונים (למשל tf) ע"פ פרמטרים שונים, לאחר בדיקות שונות הגענו למסקנה שהתוצאות הטובות ביותר מתקבלות כאשר כל מילה במסמך מנורמלת לפי max-tf (ולאו דווקא לפי אורך המסמך) וכאשר וקטור השאילתה מנורמל גם הוא לפי מספר המופעים המקסימלי של מילה בשאילתה. ביצענו ריצות רבות עבור שאילתות שונות בגישות השונות שמימשנו עם הערכים השונים ובנוסף ניסינו להוריד ולהוסיף יכולות שונות מהפרסר על מנת לבדוק האם איכות התוצאות עולה או יורדת. (מפורט בסעיף 2)
3. כדי להחליט על השיטה המיטבית מבין השיטות שמימשנו, הרצנו שאילתות רבות ועבור כל שאילתה בדקנו את הביצועים והתוצאות שחזרו מכל גישה. קראנו את הטוויטים בעצמנו כדי להבין את רמת הרלוונטיות, בנוסף השוונו מול השאילתות המדורגות שקיבלנו מה-benchmark dataset, וגם הסתכלנו בערכים שהתקבלו בטסטים מ-precision, recall ו-MAP עבור המסמכים. לבסוף שיקללנו את כל התוצאות וכך בחרנו את השיטה הטובה ביותר.
4. השתמשנו בקוד פתוח כדי לממש את הגישות לחיפוש (wordNet, Spell correction) מחבילת nltk ובמימוש spell corrector מgitHub. שני קודים פתוחים אלו מכילים מאגרי מילים שבנבדקו ואומנו והם המתאימים והמומלצים על ידי סגל הקורס בכדי לממש את הגישות הנ"ל.
5. להשלים
6. חמשת הציוצים הראשונים עבור השאילתות הנתונות:

שאילתה 1:

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

1. 

השאילתה מכילה את המילים SARS, hydroxychloroquine, paper ו-2005.



השאילתה מכילה את המילים SARS, Fauci, hydroxychloroquine, paper ו-2005.

1. יייכ



השאילתה מכילה את המילים SARS, Dr Anthony Fauci, hydroxychloroquine, paper ו-2005.

1. יככעיע

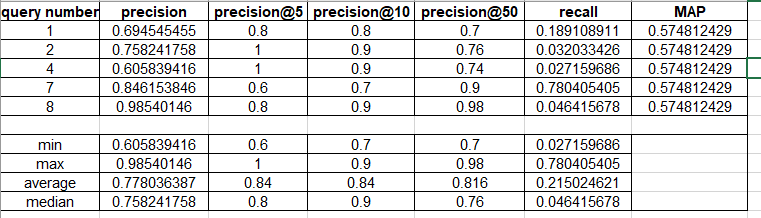
שאילתה 2:

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

1. יחי



1. יכעייעיכ
2. נציג את מדדי הביצועים של השאילתות 1,2,4,7,8 :



הצגה גרפית של כל המדדים:

1. היתרונות במנוע שלנו:

* הקוד שלנו אינו כולל שימוש בחבילות חיצוניות כבדות פרט ל-spell checker, כך שרוב הקוד מומש על ידנו ולא תלוי במקורות חיצוניים
* זמן הריצה לבניית האינדקס ולהחזרת תוצאות לשאילתה הוא מהיר מאוד.
* האינדקס שבנינו הוא יעיל ומייצג בצורה טובה את הקורפוס, הוא בנוי משלושה מילונים יחסית גדולים שמכילים כמות גדולה של מידע ולמרות זאת המנוע עובד בצורה יעילה, מהירה ומיטבית.

החסרונות במנוע שלנו:

* לדעתנו הדירוג המספרי המתקבל מתוצאות השאילתות לא מייצג בצורה הולמת את הרלוונטיות של המסמכים. מסמכים שנמצאו רלוונטיים מאוד לשאילתה בבדיקה ידנית שלנו קיבלו תוצאות לא מספקות בעינינו.
* למרות העבודה המרובה למציאת שיטות חיפוש יעילות ומגוונות (מימוש 4 שיטות) התוצאות הטובות ביותר מתקבלות בחיפוש ע"י Searcher בלבד ללא שיטות חיפוש.
* מבדיקה שנעשתה תוך כדי ניסיונות לשפר את תוצאות המדדים שלrecall, precision, MAP שמנו לב שחוזרות תוצאות רלוונטיות בדירוגים יחסית נמוכים. כלומר, כאשר מגבילים את כמות התוצאות שיורדת (מהסוף) מדד הrecall יורד בצורה משמעותית ביחס לprecision, MAP.