אוניברסיטת בן-גוריון בנגב
הפקולטה להנדסה
המחלקה להנדסת מערכות מידע
אחזור מידע 2020-2021

צוות הקורס: ד"ר ניר גרינברג
סהר בריבי, עמית ליבנה ובוריס סובול

פרויקט תכנות

בניית מנוע לאחזור מסמכים

הנחיות כלליות

מטרת הפרויקט: מטרת הפרויקט הינה ליישם את הנלמד בקורס ולהקנות ידע מעשי בפיתוח ובהערכה של מנוע לאחזור מסמכים מתוך מאגר מסמכים.
שפת הפרוייקט: הפרויקט יבוצע בשפת פייטון בלבד. עליכם לבדוק כי העבודה שלכם עובדת בסביבת העבודה במחשבי מעבדות המחלקה, כל הבדיקות יבוצעו שם, עבודה שלא תעבוד במחשבי המעבדה לא תוכל להיבדק!
הרכבי צוותים: את הפרויקט יש לבצע בזוגות בלבד, חובה להירשם כאן.
אופן ההגשה: יש לעלות את כלל הקבצים פרט לקבצי הקלט (קוד מתועד, קובץ תוצאות ודו"ח) כתיקייה מכווצת בשם Search_Engine (יש לשים לב להוראות המפורטות בקובץ instructions.txt הנמצאות בתיקייה של שלד הפרוייקט) למערכת ההגשה הייעודית תחת לשונית הקורס - חלק א'. שימו לב! המערכת נבדקה ותומכת ברגע ב Google Chrome בלבד, אנא הגישו דרכו בלבד כרגע!!!
אתר המערכת - https://subsys.ise.bgu.ac.il/submission/login.aspx
יסופקו בדיקות בסיסיות לקוד במערכת וזאת על מנת לאפשר בדיקה שלכם שהקוד אכן עובד ללא שגיאות. באחריותכם לוודא שהקוד שאתם מגישים ניתן להרצה תקינה על כל חלקיו. שימו לב גם להנחיות ההגשה שמפורטות תחת כל חלק בנפרד (הנחיות הגשה לחלק א׳ וכן הנחיות הגשה לחלק ב׳).
עבודה עם virtual environment: על מנת שתוכלו לוודא שהקוד שלכם תקין ושלא ביצעתם התקנות שאין לשרת תמיכה בהן, אנו ממליצים לכם לעבוד עם סביבת עבודה מותאמת. ניתן למצוא הנחיות להפעיל אותה ואת השלד של העבודה כאן. יש לשים לב לעקוס אחרי ההנחיות שנמצאות בקובץ (instructions.txt).
בנוסף רשימה מלאה של החבילות המותקנות בשרת הבדיקות וגרסאותיהן ניתן למצוא כאן. שימו לב כי התקנת חבילות נוספות אשר אינן נמצאות ברשימה זאת, יש לרכז ולהעביר בצורה מרוכזת.
לטובת כל מי שעובד עם pyCharm ישנו מדריך וידאו על איך להוסיף את סביבת העבודה החדשה שנוצרה לאחר הורדת הפרויקט והרצת הסקריפט מתוך ה - GitHub. ניתן לצפות בו ע"י לחיצה כאן.

פורמט הגשה: ת.ז1_ת.ז2 (לדוגמא: 234567890_123456789.zip).
שאלות והנחיות: יש לשאול שאלות על הפרויקט באמצעות הפורום במודל בלבד (שאלות הרלוונטיות לכלל הכיתה שישלחו למייל לא ייענו).
דחיות: לא יינתנו דחיות מכל סיבה שאינה מוכרת רשמית על ידי האוניברסיטה (מילואים, אשפוז וכד׳).

לכל המאחרים ללא אישור יורדו 10 נקודות על כל יום איחור בהגשה הן של הקוד והן של הדו"ח. אין להגיש את העבודה באיחור העולה על 4 ימים.
העתקות: העתקות מכל סוג שהוא - הן בין הפרויקטים השנה והן מעבודות משנים קודמות יתגלו בקלות (אנו בודקים את עבודותיכם מול עבודות של שנים קודמות ובתוכנות בדיקה וכן ידנית), העתקות יובילו את הסטודנטים לוועדת משמעת ולהשלכות הנגזרות מכך.
הפרויקט תורם מאוד להבנת הקורס ובסופו של דבר להצלחה במבחן לכן מומלץ בחום להשקיע בו ולהפיק ממנו את המרב.
שימוש בקוד פתוח, חבילות ו-API-ים חיצוניים: במידה ואתם מעוניינים להשתמש בחבילות קיימות או קוד פתוח בעל תלויות מסויימות, **עליכם לקבל את אישורו של בודק התרגילים מבעוד מועד שאכן ניתן לעשות זאת. שימוש ב-API חיצוניים גם כן טעון אישור. פנייה לשירות חיצוני בזמן חיפוש / אינדוקס עשויה להאט משמעותית את ביצועי המנוע שלכם, מה שיגרור פגיעה בציון.** בכל אופן, יש יש לציין בדוח היכן בדיוק השתמשתם בקוד חיצוני, לצרף כתובת של האתר או השירות בו השתמשתם ולהסביר כיצד השתמשתם בו.

שימו לב!! חובה להגיש מנוע עובד (כלומר לקבל ציון עובר על המנוע) על מנת לעבור את הקורס!!

שלבי הפרויקט

הפרויקט יבוצע בשלושה שלבים (יפורטו בהרחבה בהמשך) –
1. **חלק א':** (30 נק׳) בחלק זה נבנה מנוע חיפוש שלם, אך בסיסי. תדרשו לממש עיבוד ראשוני של מאגר המסמכים, בנייה של אינדקס הופכי ודירוג מסמכים.
2. **חלק ב'**: (15 נק׳) תיוג מסמכים רלוונטים ומימוש של מדדים להערכה.
3. **חלק ג׳**: (55 נק׳) מימוש שיטות מתקדמות יותר ושיפור תוצאות המנוע ובניית ממשק גרפי למשתמש.

זמני הגשה

1. **חלק א':** יוגש בתאריך 29.11.2020
2. **חלק ב'**: יוגש בתאריך 20.12.2020
3. **חלק ג׳**: יוגש בתאריך 17.01.2021

תאריכי ההגשה של הפרוייקט ומועדים שונים בקורס נמצאים בלוח שנה כאן

הערות כלליות

- המחלקות המפורטות בהמשך הינן מחלקות חובה למימוש. על מנת לממש את חלקן יש צורך בהוספת מחלקות נוספות – הרגישו חופשי להוסיף מחלקות (עם תיעוד מתאים).
  זכרו! שימוש במחלקות בצורה נכונה יקל את העבודה, יביא לניצול נכון של הזיכרון וכן יסייע בשלב ה-debugging.
- יש לעבוד בצורה מסודרת על מנת לאפשר התמצאות בקוד ועל מנת לאפשר שינויים בהמשך במידה ותידרשו לכך.

חלק א'

בחלק זה נבנה מנוע חיפוש שלם, אך בסיסי. עליכם לממש עיבוד ראשוני של מאגר המסמכים, בנייה של אינדקס הופכי ודירוג מסמכים. עומד לרשותכם מאגר מסמכים גדול של ציוצים מטוויטר בנושא קורונה (מתואר בנספח א׳), מספר שאילתות לדוגמא הנמצאות כאן, וקוד שמבצע אחזור בצורה נאיבית ביותר. עליכם לשפר ולהרחיב את הקוד הבסיסי במספר אופנים:

1. עיבוד הטקסט: עליכם להיות מסוגלים לחלץ מהטקסט URL-ים, תיוגים של חשבונות טוויטר, האשטגים ועוד ולעמוד במספר חוקים כפי שמוגדר בנספח ב׳. כמו כן, עיבוד הטקסט צריך לאפשר הרצה עם או בלי stemming.
2. בניית אינדקס המורכב ממילון וקבצי posting, כמתואר בנספח ג׳.
3. דירוג מסמכים בהתאם לשיטת שהוגדרה לכם מבין השיטות שמפורטות בנספח ד׳. עבור כל שאילתה בקובץ השאילתות שניתן לכם, עליכם להחזיר רשימה מדורגת של כ-2000 המסמכים הרלוונטיים ביותר לשאילתה. שימו לב שבחלק ג׳ של הפרויקט תדרשו לממש מספר שיטות נוספות מעבר לשיטה שהוגדרה לכם בחלק א׳, לכן מימוש של מספר שיטות כבר בחלק א׳ יזכה אתכם בניקוד נוסף.

הדאטה העומד לרשותם מתואר מטה בנספח א׳, וקוד בסיסי המשתמש בו לאחזור נמצא כאן.

לצורך חלק זה יש להוריד מאתר הקורס:

- קובץ נתונים, כמתואר בנספח א׳.
- קובץ שאילתות הנמצא כאן

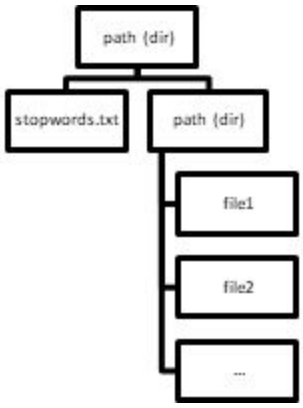
הבהרה: לפני תחילת העבודה התכנותית מומלץ לקרוא את הנתונים אותם אתם נדרשים לפרט בדו"ח ולהיערך לכך בהתאם.

מרכיבי הציון

חלק א' (30% מהציון):
5% - הערכת הקוד – הקוד צריך להיות מודולארי מתועד ועובד.
10% - יעילות הפתרון (זמן ריצה וזיכרון, יקבע באופן יחסי בהתאם לגישה שתממשו).
10% - איכות התוצאות (מספר התוצאות הרלוונטיות שהחזרתם, יקבע באופן יחסי בהתאם לגישה שתממשו)
5% - איכות ושלימות הדו"ח.

אנו נריץ בדיקות אוטומטיות על הקוד שיוגש - בעיות בהרצה של הבדיקות האוטומטיות יפגעו בציון.

עליכם לממש את המחלקות הבאות:



מחלקת ReadFile

מחלקה שתקרא את מאגר המסמכים. הסבר מפורט נמצא ב**נספח ב׳**

מחלקת Parse

מחלקה שתפרק כל ציוץ ל-terms. ה-parser צריך להתאים לציוצים בקבצים

שקיבלתם. ניתן לבצע פירסור בכל דרך אפשרית, אך **עליכם לעמוד, לכל הפחות, בחוקים המופיעים בנספח ב׳.**

מחלקת Stemmer

ניתן להשתמש ב-Porter's stemmer, ב-stemmer קוד פתוח לבחירתכם, או לממש מחלקה כזו בעצמכם.

מחלקת Indexer

ה-Indexer מקבל את המילים מה- parser ובונה את ה- inverted index. **ההנחיות לבניית האינדקס מופיעות בנספח ג׳.**

מחלקת Searcher

תפקידה לבצע את השאילתות.

מחלקת Ranker

תפקידה לדרג את התשובות לשאילתות על פי נוסחת דירוג שאתם תפתחו.

הסברים על ה־Searcher, הRanker ופירוט השיטות שעליכם לממש נמצאים בנספח ד׳.

הנחיות הגשה

כפי שצוין קודם לכן, יש להגיש את הפרוייקט כתיקייה מכווצת בשם Search_Engine (יש לשים לב להוראות המפורטות בקובץ instructions.txt הנמצאות בתיקייה של שלד הפרוייקט) למערכת ההגשה הייעודית באתר הזה תחת לשונית הקורס - חלק א'.
בתיקייה יש לשים את הקבצים הבאים:

1. חבילת הפרויקט (קוד מקור מתועד).
2. קובץ הוראות הפעלה (Readme).
3. דו"ח - קובץ Word (קובץ אחר לא ייבדק!). הסבר מפורט על הדו״ח נמצא בנספח ו׳.
4. קובץ csv עם תוצאות הדירוג שיצרתם. הקובץ צריך לכלול את השדות הבאים:
    a. Query_num - מספר השאילתא עבורה הטוויטים רלוונטים
    b. Tweet_id - מספרי הציוצים הרלוונטים
    c. Rank - הדירוג של הציוץ ביחס לשאילתא

חלק ב׳

בחלק זה של הפרויקט אתם נדרשים

1. לממש מספר מדדי הערכה לתוצאות אחזור כמפורט מטה

2. לבצע תיוג **ידני** של כ-300 זוגות של ציוץ + שאילתה כרלוונטים או לא. כל אחד ואחת מכם יקבל אוסף (ייחודי) של ציוצים + שאילתות לתיוג, ועליכם למלא עמודה נוספת ובה התיוגים שלכם. תיוגים שתייצרו ישמשו אותנו לבניית ה-ground truth עבור חלק ג׳ של הפרויקט.

מדדי הערכה אותם תצטרכו לממש הם:

- Precision - אפשרות לחישוב עבור שאילתא בודדת וכן עבור למנוע כולו.
- Recall - אפשרות לחישוב עבור שאילתא בודדת וכן עבור למנוע כולו.
- Precision@N - חישוב הערך עבור N טוויטים (5,10,15,30,50).
- MAP למנוע.

מרכיבי הציון

חלק ב' (15% מהציון):
10% - נכונות המדדים שמימשתם
3% - ביצוע התיוגים
2% - איכות התיוגים (ימדד ע״י inter-coder reliability)

הנחיות הגשה

הנחיות מפורטות יפורסמו לאחר הגשת חלק א׳.

חלק ג׳

בחלק זה עליכם להוסיף חלקים נוספים למנוע שבניתם בחלק א׳, וכן לבדוק ולשפר את ביצועי המנוע.

מרכיבי הציון

חלק ג' (55% מהציון):
10% - יעילות הפתרון (זמן ריצה וזיכרון, יקבע באופן יחסי).
17% - יצירתיות הפתרון (נלקחת בחשבון גם יצירתיות שלא הניבה בהכרח שיפורים).
18% - איכות התוצאות (מרכיב זה בציון ייקבע באופן יחסי לשאר הפרויקטים בכיתה).
10% - איכות ושלימות הדו"ח.

הנחיות מפורטות לגבי הדרישות בחלק ג׳ יפורסמו בהמשך.

אנא וודאו לפני הגשה: שהגשתם את הדוח במלואו, שהקוד מתועד, עובד ולא נופל, זרקו שגיאות בהתאם. וודאו שהקוד עובד במעבדה. אי מילוי של הוראות אלו יפגע בציונכם בצורה משמעותית.

בהצלחה!

נספח א' - הסבר על הדאטה

קובץ הנתונים שתקבלו הוא קובץ Data.zip המכיל מספר קבצים כאשר כל אחד מהם הוגדר כקובץ parquet ומכיל טוויטים. קובץ הנתונים נמצא כאן. מיפוי השדות הוא כדלקמן:

- tweet_id - מספר מזהה של הציוץ
- full_text - הטקסט שמופיע בציוץ
- urls - מיפוי בין הurl המקוצר המוצג בטוויט לבין הurl המורחב האמיתי. אם קיים URL בציוץ הוא יופיע כאן, אחרת יהיה ריק.
- indices - המיקום של הurls בתוך הטקסט
- retweet_text - טקסט של ה-retweet במידה וקיים
- retweet_urls -מיפוי בין הurl המקוצר המוצג בטוויט לבין הurl המורחב האמיתי. אם קיים URL בretweet הוא יופיע כאן, אחרת יהיה ריק.
- retweet_indices - המיקום של הurls בתוך הטקסט של הretweet
- quoted_text - טקסט של ה-quote במידה וקיים
- quoted_urls - מיפוי בין הurl המקוצר המוצג בטוויט לבין הurl המורחב האמיתי. אם קיים URL בquote הוא יופיע כאן, אחרת יהיה ריק.
- quoted_indices - המיקום של הurls בתוך הטקסט של הquote
- retweet_quoted_text - טקסט של ה-retweet במידה והוא גם quoted.
- retweet_quoted_urls - מיפוי בין הurl המקוצר המוצג בטוויט לבין הurl המורחב האמיתי. אם קיים URL הוא יופיע כאן, אחרת יהיה ריק.
- retweet_quoted_indices - המיקום של הurls בתוך הטקסט של הretweet_quoted

מאגר הטוויטים נראה כך:

tweet_id	tweet_date	full_text	urls	indices	retweet_text	retweet_urls	retweet_indices	quoted_text	quote_urls	quoted_indices	retweet_quoted_text	retweet_quoted_urls	retweet_quoted_indices
1291795943227043847	Fri Aug 07 17:58:47 +0000 2020	RT @Charlotte3003G: Spoke to a wonderful respiratory consultant I've know for over	{}	[]	Spoke to a wonderful respiratory consultant I've know for over 20 years. I asked her about mask wearing. Her respon… https://t.co/w57gzbWk67	{\https://t.co/w57gzbWk67\":\"https://twitter.com/i/web/status/1291767035203072000\"}"	[[117,140]]						

		20 years. I asked her about mask wearing. Her response:...											
1291795945529778176	Fri Aug 07 17:58:48 +0000 2020	RT @xvniji a: wearing a mask and being soft spoken do not mix	{}	[]	wearing a mask and being soft spoken do not mix	{}	[]						
1291795983509213185	Fri Aug 07 17:58:57 +0000 2020	POST-COVID 19 https://t.co/3gAU9CpG0P	{\https://t.co/3gAU9CpG0P\":\"https://www.bbc.co.uk/news/resources/idt-dc2d6e2d-3ab4-42de-8d03-bb7eda5fff8e\"}"	[[15,38]]									

נספח ב׳ - הנחיות ל Reader & Parser

מחלקת ReadFile

מחלקה שתקרא את מאגר המסמכים. המחלקה תדע לקבל path של תיקייה בה יושבים כלל הקבצים (אחרי ביצוע unzip). **אין צורך** לבצע unzip בקוד, הניחו שהתיקייה תהיה אחרי ביצוע ה zip. הקבצים יהיו קבצי parquet שיהיה עליכם לקרוא.

מחלקת Parse

מחלקה שתפרק כל ציוץ ל-terms. ה-parser צריך להתאים לציוצים בקבצים שקיבלתם. ניתן לבצע פירסור בכל דרך אפשרית, אך **עליכם לעמוד, לכל הפחות, בחוקים המפורטים מטה.** במידה ויש מספר חוקים המתנגשים האחד בשני אתם רשאים לפעול לפי שיקול דעתכם (לדוגמא, לשמור את ה-term בשתי דרכים אפשריות).

בנוסף לחוקים המופיעים מטה, עליכם:

1. להגדיר שני חוקים משלכם: להסביר את ההיגיון שלהם ולממש אותם. **בדו"ח יש להדגים כיצד החוקים באים לידי ביטוי בשני ציוצים שונים ב-Dataset**.
2. אין צורך להתייחס באופן מיוחד לסימני הפיסוק (אלא אם כן הם מהווים חלק מכלל), הם יכולים לשמש אתכם על מנת להפריד בין המילים/המשפטים.
3. יש להוריד stop-words על פי הרשימה שפורסמה באתר. יש לשים לב כי ה-STOP WORD אינה סותרת את הכללים המופיעים מטה (לדוגמא: THE DOLLAR). במקרה של סתירה, קרי מילה המופיעה ברשימת ה-stop-words הינה חלק מ-term כפי שהוגדר בכללים מטה, הרי שהמילה איננה stop-word ולכן אין לנפות אותה. את קובץ ה-stop-words יש לשים באותו המיקום של מאגר המסמכים. שימו לב שאתם יכולים להרחיב את רשימת הstop-words במידת הצורך.
4. יש לאפשר ביצוע stemming.

להלן החוקים לפירוק הטקסט מהציוצים ל-terms:

<u>טיפול בHashtags</u>

כל טקסט שמתחיל עם הסימן # נחשב כהאשטאג. נרצה להפריד את ההשטגים למילים נפרדות ולשמור כל מילה בנפרד. יש לשמור את ההאשטגים הן כמונח אחד, והן כמילים נפרדות. לדוגמא:

מופיע במסמך	יש לשמור כך
#stayAtHome	stay, at, home, #stayathome

stay, at, home, #stayathome	#stay_at_home

URL-ים

יש להפריד את הURL-ים למילים נפרדות, כאשר את הדומיין עליכם לשמור כמונח שלם. כלומר, הURL הבא:
https://www.instagram.com/p/CD7fAPWs3WM/?igshid=o9kf0ugp1l8x
יופרד למילים: https, www, instagram.com, p, CD7fAPWs3WM, igshid, o9kf0ugp1l8x

שימו לב, ניתן להוסיף לרשימת הstopwords הקיימת מילים נוספות שלדעתכם רלוונטיות לדאטה הנוכחי.

תיוגים

טקסט שמתחיל ב@ הוא תיוג. נרצה לשמור את התיוגים עם הסימן שלהם.

אותיות גדולות/קטנות

מילים שהאות הראשונה שלהם היא תמיד אות גדולה, בכל הקורפוס, ישמרו עם אותיות גדולות בלבד. מאידך, אם מילה מופיע לעיתים עם אות גדולה ולפעמים ללא אות גדולה נשמור אותה עם אותיות קטנות בלבד.

דוגמא 1, המשפטים הבאים מופיעים כך בקורפוס:

Sentence #1: “First, …. “

Sentence #2: “At first, we …”

במקרה הנ"ל יש לשמור את המילה first בצורה של אותיות קטנות בלבד.

דוגמא 2, המשפטים הבאים מופיעים כך בקורפוס:

Sentence #1: “NBA ….”

Sentence #2: “GSW is the NBA champions”

במקרה הנ"ל יש לשמור את המילה NBA בצורה של אותיות גדולות.

דוגמא 3, המשפטים הבאים מופיעים כך בקורפוס:

Sentence #1: “Max ….”

Sentence #2: “Max and Roy are good friends”

במקרה הנ"ל יש לשמור את המילה Max בצורה של אותיות גדולות: MAX.

אחוזים

כל מספר אשר מצורף אליו אחוז בכל אחד מהפורמטים הבאים ישמר כ:
NUMBER %

1. Number% (e.g. 6%)
2. Number percent (e.g. 10.6 percent)
3. Number percentage (e.g. 10.6 percentage)

דוגמאות:

מופע בציוצים	יש לשמור כך
6%	6%
10.6 percent	10.6%
10.6 percentage	10.6%

מספרים ללא יחידות

מדובר במספרים ללא סימון נוסף הצמוד אליהם כמו דולר או אחוז (חוקים לגבי מספרים עם יחידות מופיעים בהמשך).

יש לשמור מספרים לא שלמים עם דיוק של עד 3 ספרות אחרי הנקודה העשרונית.

יש להתייחס למספרים בצורה הבאה:

1. **כל מספר שהוא מעל אלף** (1,000) יש לייצג בצורה של NUMBER K/M/B.
   a. כל מספר בין אלף (כולל) למיליון (לא כולל) ישמר עם K. לדוגמא:

מופיע במסמך	יש לשמור כך
10,123	10.123K
123 Thousand	123.456K
1010.56	1.01K

   b. כל מספר בין מיליון (כולל) למיליארד (לא כולל) ישמר עם M. לדוגמא:

מופיע במסמך	יש לשמור כך
10,123,000	10.123M
55 Million	55M

c. כל מספר מעל למיליארד ישמר עם B. לדוגמא:

מופיע במסמך	יש לשמור כך
10,123,000,000	10.123B
55 Billion	55B

2. **כל מספר שהוא מתחת לאלף -** מספרים בצורות השונות ישמרו כפי שהם, לדוגמא מספר 204 ישמר כ- 204 (אין לפרק את המספר אלא להשאירו בשלמותו, כלומר לא לפרק ל2,0,4), מספר עשרוני גם יישמר כפי שהוא, לדוגמא 35.66 ישמר כ-35.66. במידה ויש מספר שאחריו מגיע שבר לדוגמא: 35 3/4 יש לשמור את המספר כולל השבר.

<u>שמות וישויות</u>

יש לשמור באינדקס שמות של ישויות המופיעות בטקסט בשני טוויטים או יותר. באופן הבסיסי ביותר, ניתן לזהות ישויות כרצף של terms המורכב ממילים שמתחילות באות גדולה, אשר מופיע בשני טוויטים שונים במאגר או יותר. ניתן גם לזהות ישויות בדרכים מורכבות יותר באמצעות הפעלת Part-of-speech tagger על הטקסט ולאחריו named-entity recognition. דוגמא לישות: Alexandria Ocasio-Cortez. גם כאן ניתן לשמור באינדקס את ה-terms ממנו מורכב השם כל אחד בנפרד וכולם ביחד.

נספח ג' - הנחיות ל-Indexer

ה-Indexer מקבל את המילים מה- parser ובונה את ה- inverted index.
ה- inverted index כולל (חיזרו על ההרצאה והתרגול כדי להבין מה כל מבנה מציין):
1. **מילון** - מילון שיועלה לזיכרון הראשי בחיפוש. המילון ימומש במבנה כראות עיניכם.

2. **קבצי Posting** - יש לבנות קבצי Posting שיאוחסנו בדיסק ויכללו מידע על כל ה-terms והמסמכים במאגר לפי בחירתכם.

מספר הנחיות לבניית האינדקס:

1. **אין לשמור את קובץ ה-posting באמצעות DB!** (גם לא בקובץ CSV), יש לשמור את הנתונים באמצעות קבצים פשוטים (לדוגמא: קבצי txt) בדיסק הקשיח.
2. בעת ביצוע תהליך ה-indexing **אין להחזיק את כל המידע על ה-terms בזיכרון הראשי ואז לכתוב אותם במרוכז לקובץ ה-posting וכך ליצור את המילון**. כלומר יש ליישם או לפתח שיטה שבונה את קובץ ה-Posting באופן הדרגתי, כלומר מוסיפה לו כל פעם עדכונים לקבוצת טוויטים חלקית של המאגר, תוך כדי העלאה לזיכרון של חלקים מה-Posting. גודל הקבוצה החלקית של הטוויטים שיטופלו בכל שלב נתון לשיקול דעתכם. עליכם לציין בדו"ח את הסיבות לבחירת גודל קבוצת הטוויטים החלקית וכן להציג תיעוד תהליך יצירת הקבצים ההופכיים (המילון וקובץ posting). כמו כן בעת ההגנה הסופית על המנוע יהיה עליכם להציג ולהסביר את קטעי הקוד הללו.
3. יש ליישם אלגוריתם יעיל לבניית- inverted index.
4. מומלץ להשקיע מחשבה במבנה בו תשמרו את קבצי ה-posting – החלוקה לקבצים השונים, הצורה בה תשמרו את הנתונים השונים על כל term או מסמך וכדומה. מבנה זה ישפיע על מהירות יצירת ה-inverted index, מהירות האחזור וכן נפח האחסון הנדרש.
5. עבור כל term עליכם לשמור לפחות את:
   a. כמות הטוויטים בהם הוא מופיע (df).
   b. כמות הפעמים בהם הוא הופיע בכל טוויט (tf).
   c. רשימה של הטוויטים בהם הוא מופיע.
6. עבור כל מסמך עליכם לשמור לפחות את:
   a. תדירות ה-term הנפוץ ביותר (max_tf).
   b. כמות המילים הייחודיות במסמך.
7. עליכם לשמור **לפחות 2 פריטי אינפורמציה נוספים על ה-terms או הטוויט.** אינפורמציה זו תוכל לעזור לכם בעתיד בחלק ב' של העבודה כאשר תדרשו לאחזר מסמכים רלוונטיים לשאילתות.

נספח ד' - Searcher & Ranker

מחלקת Searcher

תפקידה לבצע את השאילתות. המחלקה תקבל שאילתה (מילה או אוסף של מילים עם רווחים ביניהם), המחלקה תנתח את השאילתה בהתאם לניתוח הטקסט שנעשה על המסמכים ותחזיר את הטוויטים הרלוונטיים ביותר לשאילתה באופן מדורג (באמצעות שימוש במחלקת Ranker המפורטת בהמשך).

- מספר הטוויטים המוחזרים לשאילתה מוגבל ב-2000 (כלומר יש להחזיר את 2000 טוויטים הרלוונטיים ביותר לפי הדירוג רלוונטיות).

מחלקת Ranker

תפקידה לדרג את התשובות לשאילתות על פי נוסחת דירוג שאתם תפתחו. בחלק זה אתם רשאים להשתמש בכל אינפורמציה ששמרתם בinverted index או מהמסמך שאתם מוצאים לנכון.

בחלק זה יהיה עליכם לממש אחת מהשיטות המפורטות בנספח ד'. שיטות אלו יעזרו לכם למצוא את הדמיון בין השאילתא לבין הטוויטים.

1. Word2vec
2. GloVe
3. SVD (or Latent Dirichlet Allocation
4. WordNet
5. Thesaurus
6. Global Method
7. Local Method
8. Advanced parser
9. Spelling correction

נספח ה' - דרישות לדו"ח בחלק א'

הדוח צריך להכיל:

1. עיצוב התוכנה:
   a. הסבר מפורט על אופן פעולת התכנית שבניתם, והסבר על הגישות שמיממשתם בפרוייקט, נדרשות ונוספות. אין צורך לפרט על כל הפונקציות שמופיעות בכל מחלקה, אלא להסביר באופן כללי על המחלקות והחלקים שמימשתם.

b. יש להסביר על האופן שבו התמודדתם עם מגבלת הזיכרון של המחשב והפעולות שנקטתם על מנת להביא לזמן ריצה מיטבי.
c. יש להסביר באיזה אופן שמרתם את קבצי ה-Posting, סוג הקבצים, כמות הקבצים, מה מכיל כל קובץ וכדומה. יש לנמק את הבחירה.
d. עליכם לציין את הסיבות לבחירת גודל קבוצת הטוויטים החלקית וכן להציג תיעוד תהליך יצירת הקבצים ההופכיים (מילון וקובץ posting).
e. עליכם לציין את פריטי האינפורמציה הנוספים ששמרתם ולהסביר מדוע שמרתם דווקא אותם.
f. עליכם להסביר מה שני החוקים שהוספתם, יש להדגים כיצד החוקים באים לידי ביטוי בשני טוויטים שונים ב-Dataset.
g. עבור כללים/ חוקים נוספים שהוספתם (ב-Parser וב-Indexer) יש להסביר את הרעיון של כל חוק וכיצד מימשתם אותו.
h. הסבירו בצורה מפורטת את האלגוריתם בו השתמשתם במחלקת Ranker
i. עליכם לציין האם השתמשתם במהלך העבודה בקוד פתוח לפרט את השירות, כתובת, היכן השתמשתם, כיצד השתמשתם.
j. כמו כן, עליכם להוסיף כל מידע נוסף שלדעתכם חשוב להבנת התכנית ע"י הבוחן.
k. מנו שלושה יתרונות ושלושה חסרונות של המנוע של המנוע שלכם על פני מנועים אחרים.

2. לאחר עיבוד המאגר יש להגיש במסמך את רשימת הפלטים הבאים (אין צורך לממש בקוד במנוע, מלבד מה שכבר נדרשתם):

a. כמה terms שונים יש במאגר לפני stemming?
b. כמה terms שונים יש במאגר אחרי stemming?
c. כמה terms שונים שהם מספרים יש במאגר?
d. הדפיסו את רשימת 10 ה-terms השכיחים ביותר במאגר לפי סדר שכיחות, ואת רשימת 10 ה-terms הכי פחות שכיחים במאגר (לפני stemming). השכיחות הינה כמות המופעים הכוללת של term במאגר (לא כמות הטוויטים בהם מופיע ה-term שזה כאמור נתון ה-df).
e. הציגו את המילים הייחודיות במאגר על גרף לפי Zipf's Law (ציר Y מבטא שכיחות של המילה במאגר כולו). תזכורת- המילה שמופיעה הכי הרבה פעמים במאגר תופיע ראשונה על ציר הX וכך הלאה (אין צורך לרשום על ציר הX את המילים עצמן). הסבירו האם העקומה שיצאה לכם אכן דומה ל Zipf's Law.
f. הציגו את גודל ה- Posting – נפח האחסון הנדרש עבור קבצי ה- Posting (ב-KB) עבור stemming וללא.
g. הציגו את משך הזמן שלקח למנוע לבנות את האינדקס על קבצי ה Corpus.

h. עבור שאילתות מספר 1,2,4,7,8 בחנו את חמשת הציוצים הראשונים שאוחזרו והסבירו למה, בהתאם לחוקים והשיטות שמימשתם, דווקא ציוצים אלו דורגו כציוצים הראשונים שאוחזרו.