אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

אוניברסיטת בן-גוריון בנגב
הפקולטה להנדסה
המחלקה להנדסת מערכות מידע
אחזור מידע 2019-2020

צוות הקורס : ד"ר ניר גרינברג
רועי דור , מקס שפירא ויפים פייטרברג

פרויקט תכנות – חלק א׳

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 הפרויקט: הפרויקט יבוצע בשפת Java 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.
אופן ההגשה: יש לעלות את כלל הקבצים (קוד מתועד, הוראות הפעלה ודו"ח) כתיקייה מכווצת לשרת ה-FTP של המחלקה (תיקייה להעלאת הפרויקטים תיפתח ותפורסם בהמשך).
בנוסף, יש להעלות את קבצי הקוד למערכת ההגשה הייעודית תחת לשונית הקורס - חלק א'.
אתר המערכת - https://subsys.ise.bgu.ac.il/submission/login.aspx
יסופקו בדיקות בסיסיות לקוד במערכת וזאת על מנת לאפשר בדיקה שלכם כי הקוד אכן עובד ללא שגיאות, באחריותכם לודא כי הקוד שאתם מגישים ניתן להרצה תקינה על כל חלקיו.
פורמט הגשה: ת.ז1_ת.ז2 (לדוגמא: zip.123456789_234567890).
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.

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

שלבי הפרויקט

הפרויקט יבוצע בשני שלבים (יפורטו בהרחבה בהמשך) –

1. **חלק א'**– בחלק זה תבצעו עיבוד מקדים למאגר (קריאת המסמכים מהקבצים, פירסור, פעולות על טקסט, בניית ה-inverted index - מילון וקבצי posting).
2. **חלק ב'** – בחלק זה תפתחו את המנוע המאפשר חיפוש מסמכים העונים על שאילתה ודירוגם, ובניית ממשק גרפי למשתמש.

מספר הבהרות:

- יש לצרף לחלקים א' וב' דו"ח כמפורט בהמשך. שלימות ואיכות הדו"ח מהווים חלק מהציון על העבודה.
- 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.
- במידה ואתם מעוניינים להשתמש בקוד פתוח באחד מחלקי העבודה אתם רשאים, יש לציין בדוח היכן השתמשתם בקוד פתוח, לצרף כתובת של האתר או השירות בו השתמשתם ולהסביר כיצד השתמשתם.
- לאחר הגשת המנוע השלם (חלק ב'), תתבצע בדיקה פרונטאלית. בבדיקה תידרשו:
  - להציג מנוע **עובד** ולהסביר את אופן עבודתו.
  - להציג את יכולת המנוע לענות על שאילתות קיימות חדשות שתוצגנה למנוע.
  - **יש להגיע בהרכב מלא של הזוג לבחינה, ייתכן ציון שונה לסטודנטים בזוג על פי התרשמות הבוחן.**
  - **אי אפשר לקבל עובר על המנוע ללא בחינה פרונטאלית.**

מומלץ בחום לקרוא את ההנחיות בשלמותן (חלק א' וב') בטרם תחילת ביצוע הפרויקט (גם לפני חלק א'). למבני הנתונים, הקבצים והנתונים עצמם בהם תבחרו להשתמש בחלק א' ישנה השפעה רבה על קלות המימוש של חלק ב' וכמובן על טיב התוצאות במענה על שאילתות בחלק ב'. יובהר כי **הנכם רשאים לבצע שינויים בחלקי קוד הקשורים לחלק א' בעת העבודה על חלק ב'.** במידה ויש לכם רעיונות חדשים, נוכחתם לדעת כי חסרים לכם נתונים כאלה ואחרים או על סמך ההערות שקיבלתם במסגרת המשוב על חלק א'. את השינויים שביצעתם במחלקות מחלק א' עליכם לציין בדו"ח של חלק ב'.

מרכיבי הציון

חלק א' (40% מהציון):

- 50% - הערכת הקוד – הקוד צריך להיות מודולארי מתועד ועובד.
- 20% - יעילות הפתרון (זמן ריצה וזיכרון).
- 30% - איכות ושלימות הדו"ח.

חלק ב' (40% מהציון):

- 30% - הערכת הקוד – הקוד צריך להיות מודולארי מתועד ועובד (**ללא קוד עובד לא ניתן לעבור את הקורס!**).

- 10% - יעילות הפתרון (זמן ריצה וזיכרון).
- 20% - יצירתיות הפתרון (נלקחת בחשבון גם יצירתיות שלא הניבה בהכרח שיפורים).
- 20% - איכות התוצאות (מרכיב זה בציון ייקבע באופן יחסי לשאר הפרויקטים בכיתה).
- 20% - איכות ושלימות הדו"ח.

בדיקה פרונטאלית (20% מהציון) :
- 40% - מענה על שאילתות קיימות וחדשות.
- 60% - התרשמות הבוחן מידיעות הסטודנטים ומאיכות והיצירתיות שבעבודה.

לו"ז

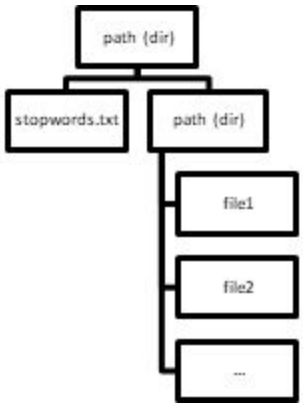
תאריך אחרון להגשת חלק א' של הפרויקט: 15/12/2019
תאריך אחרון להגשת חלק ב' של הפרויקט: 19/1/2020
תאריכי הבחינה הפרונטלית: ה-21 וה-23 לינואר 2020

חלק א'

לצורך חלק זה יש להוריד מאתר הקורס:
- מאגר מסמכים – קובץ corpus.zip המכיל תיקיות עם קבצים רבים. כל קובץ מכיל מסמכים רבים. להלן מיפוי התגיות המרכזיות:
  - <DOC> מסמל התחלה של מסמך.
  - <DOCNO> מסמל את המזהה של המסמך.
  - <DATE1> מסמל את תאריך הפרסום של המסמך בפורמט day month year.
  - <TI> מסמל את כותרת המסמך.
  - <TEXT> מסמל את תוכן המסמך.
- רשימת stop words

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

עליכם לממש את המחלקות הבאות:



מחלקת ReadFile

מחלקה שתקרא את מאגר המסמכים. המחלקה תדע לקבל path של תיקייה בה יושבים כלל הקבצים (אחרי ביצוע unzip). בכל קובץ יש הרבה מסמכים, נדרש לזהות את ההתחלה של כל מסמך ולהפריד את המסמכים בהתאם. **אין צורך** לבצע unzip בקוד, הניחו שהתיקייה תהיה אחרי ביצוע ה zip.

מחלקת Parse

מחלקה שתפרק כל מסמך ל-terms. ה-parser צריך להתאים למסמכים במאגר המסמכים. מספיק להתייחס בשלב זה רק לטקסט שנמצא בין התגים <TEXT> (בהמשך יהיה ניתן להתייחס גם לתגים נוספים). ניתן לבצע פירסור בכל דרך אפשרית, אך **עליכם לעמוד, לכל הפחות, בחוקים המופיעים מטה לגבי העיבוד של מספרים, אותיות קטנות וגדולות, וכד'**. 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 מסמכים 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.
4. יש לאפשר ביצוע stemming.

להלן החוקים לפירוק הטקסט ממסמכים ל-terms:

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: 3/4 35 יש לשמור את המספר כולל השבר.

<u>אותיות גדולות/קטנות</u>

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.

<u>אחוזים</u>

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 במסמכים	יש לשמור כך
6%	6%
10.6 percent	10.6%
10.6 percentage	10.6%

<u>מחירים</u>

יש להתייחס למחירים בצורה הבאה:

1. **כל מחיר שהוא מתחת מיליון דולר** יש לייצג בצורה של NUMBER Dollars. עליכם לכסות את הפורמטים הבאים:

   i. Price Dollars
   ii. Price fraction Dollars
   iii. $price

דוגמאות:

מופע במסמכים	**יש לשמור כך**
1.732	1.732 Dollars
22 3/4 Dollars	22 3/4 Dollars
$450,000	450,000 Dollars

2. **כל מחיר שהוא מעל מיליון דולר** יש לייצג בצורה של NUMBER M Dollars. עליכם לכסות את הפורמטים הבאים:

   i. Price Dollars
   ii. $price
   iii. $price million
   iv. $price billion
   v. Price m Dollars
   vi. Price bn Dollars
   vii. Price billion U.S. dollars
   viii. Price million U.S. dollars
   ix. Price trillion U.S. dollars

דוגמאות:

מופע במסמכים	**יש לשמור כך**
1,000,000 Dollars	1 M Dollars
$450,000,000	450 M Dollars
$100 million	100 M Dollars

20.6 M Dollars	20.6m Dollars
100000 M Dollars	$100 billion
100000 M Dollars	100bn Dollars
100000 M Dollars	100 billion U.S. dollars
320 M Dollars	320 million U.S. dollars

<u>תאריכים</u>

עליכם לשמור כל תאריך לפי הפורמט המבוקש (אין צורך למצוא פורמטים נוספים מלבד מה שנתבקשתם):

1. תאריכים שמופיעים בפורמט הבא ישמרו כ: MM-DD
   i. DD Month
   ii. Month DD

דוגמאות:

מופע במסמכים	יש לשמור כך
14 MAY, 14 May	05-14
June 4, JUNE 4	06-04

2. תאריכים שמופיעים בפורמט הבא ישמרו כ: YYYY-MM
   i. DD Month

דוגמא:

מופע במסמכים	יש לשמור כך
May 1994, MAY 1994	1994-05

בכל המקרים שמות החודשים יכולים להופיע גם באופן מקוצר בכל המקרים (3 אותיות ראשונות, לדוג': Jan).

<u>טווחים/ביטויים עם מקף</u>

טווחים וביטויים בצורות הבאות יישמרו כ-term יחיד (number יכול להיות מספר מכל סוג שהוגדר בסעיף א', word הינה מילה שאיננה מספר) :

1. Word-word (for example: Value-added)
2. Word-word-word (for example: step-by-step)
3. Number-word or Word-Number (for example: 10-part)
4. Number-number (for example: 6-7)
5. Between number and number (for example: between 18 and 24)

ניתן לשמור את הטווחים **גם** כמספרים נפרדים, לדוגמא בעבור 6-7 נשמור שלושה terms שונים 6, 7 וגם את 6-7.

<u>שמות וישויות</u>

יש לשמור באינדקס שמות של ישויות המופיעות בטקסט בשני מסמכים או יותר. באופן הבסיסי ביותר, ניתן לזהות ישויות כרצף של terms המורכב ממילים שמתחילות באות גדולה, אשר מופיע בשני מסמכים שונים במאגר המסמכים 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.

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.
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 מסמכים חלקית של המאגר, תוך כדי העלאה לזיכרון של חלקים מה-Posting. גודל הקבוצה החלקית של המסמכים שיטופלו בכל שלב נתון לשיקול דעתכם. עליכם לציין בדו"ח את הסיבות לבחירת גודל קבוצת המסמכים 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 המסמכים בהם הוא מופיע (df).
   b. כמות הפעמים בהם הוא הופיע בכל מסמך (tf).
   c. רשימה של המסמכים 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 המסמך**. אינפורמציה זו תוכל לעזור לכם בעתיד בחלק ב' של העבודה כאשר תדרשו לאחזר מסמכים רלוונטיים לשאילתות.

מחלקת GUI

ממשק למשתמש

עליכם לממש ממשק עם האפשרויות הבאות:

1. **כפתור הפעלה -** אפשרות הפעלה של התכנית על data set נבחר המכיל קבצים באותו פורמט של הקבצים שיש ב data set המופיע באתר. על הממשק להציג 2 תיבות טקסט (text box):
   - תיבה ראשונה להזנת ה-path בו נמצאת התיקייה עם מאגר המסמכים וכן רשימת ה-stop-words. יש לאפשר בחירה של תיקייה באמצעות כפתור browse.
   - תיבה שנייה להזנת ה-path בו יישמר קובץ/קבצי ה-posting והמילון בדיסק הקשיח. יש לאפשר בחירה של תיקייה באמצעות כפתור browse. (אפשרות זו נועדה לטעינה מהירה של שני האינדקסים עבור בדיקות של המנוע).
2. **Checkbox** שמאפשר **לבצע/לא לבצע Stemming**. שימו לב, יש לאפשר בנייה של המאגר בשתי הדרכים אחת אחרי השנייה – יש לשמור קובץ/קבצי posting עבור כל אחת מהאופציות. כלומר, שיהיה ניתן להריץ לדוגמא בתחילה עם stemming ולאחר מכן בלי וייווצרו קבצים שונים לשני המצבים (שלא תהיה דריסה של הקבצים).
3. **כפתור איפוס –** אפשרות לאיפוס התהליך. לחיצה על כפתור זה תביא למחיקת כלל קבצי ה-posting והמילון שנשמרו (יש להניח כי התיקייה שצוינה בתיבת הטקסט לשמירת קבצי ה-posting לא השתנתה) וכן לאיפוס הזיכרון הראשי בתוכנית.
4. **כפתור להצגת המילון** – בעת לחיצה על כפתור זה, ייפתח חלון המציג את המילון ממוין בצורה עולה (מספרים..A..B..C..). אנא הציגו את המילים וכמות המופעים הכולל שלהם במאגר בלבד.
5. **כפתור לטעינת המילון לזיכרון** – הכפתור ישתמש ב path בו נשמרים קובץ/קבצי ה-posting בדיסק הקשיח ויטען את המילון לזיכרון. כפתור זה ישמש לשלב הבדיקה הפרונטלית ויאפשר להריץ את המנוע ללא ביצוע תהליך בניית הinverted index ורק על סמך כלל קבצי ה-index (הכוללים קבצי posting והמילון) שיוגשו טרם הבדיקה. קובץ המילון שייטען לזיכרון (עם stem או בלי) ייטען עפ"י הסימון בכפתור Stem שצוין קודם לכן.

הערות

- יש לוודא כי בעת לחיצה על כפתור הפעלה מוזנים ערכים בתיבות הטקסט. במידה ולא, המשתמש יקבל הודעה מתאימה.
- בעת סיום תהליך יצירת כל אינדקס בנפרד תעלה הודעה המכילה את הנתונים הבאים:
  - מס' מסמכים שאונדקסו.
  - כמות ה-terms הייחודיים שזוהו במאגר.
  - זמן ריצה כולל של התהליך בשניות (קריאת מסמכים, פרסור ואינדוקס של כל המאגר): מרגע הפעלת התוכנית ועד לסיום בניית ה inverted index (יצירת המילון וקבצי posting).
- ניתן להניח כי מאגר המסמכים הינו סטטי. על מנת לשנות את מאגר המסמכים המשתמש יבצע איפוס ולאחר מכן יטען מאגר (תיקייה) חדש באותו הפורמט של המסמכים שניתנו לכם.
- הערה: העיצוב של הממשק צריך להיות בסיסי ועובד. אין טעם להשקיע בעיצוב כי לא ניתנות נקודות נוספות על עיצוב משופר.

הנחיות הגשה

כפי שצוין קודם לכן, יש להגיש את הדו"ח לספריית ה-ftp (הכתובת תפורסם בהמשך) את החלק הראשון של הפרויקט כתיקייה אחת מכווצת.
בתיקייה יש לשים את הקבצים הבאים:

1. חבילת הפרויקט (קוד מקור מתועד) כולל קובץ מוכן להרצה.
2. קובץ הוראות הפעלה (Readme).
3. דו"ח - קובץ Word (קובץ אחר לא ייבדק!).

בנוסף, את קבצי הקוד יש להגיש למערכת ההגשות המחלקתית בכתובת המופיעה מעלה.

הדוח צריך להכיל:

1. עיצוב התוכנה:
   a. הסבר מפורט על אופן פעולת התכנית שבניתם, אילו מחלקות יצרתם (אם הוספתם מחלקות, יש להסביר מה מטרתן ואיך הן פועלות) ותיאור השיטות של כל מחלקה.
   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 המסמכים 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 מסמכים שונים ב-Dataset.

g. עבור כללים/ חוקים נוספים שהוספתם (ב-Parser וב-Indexer) יש להסביר את הרעיון של כל חוק וכיצד מימשתם אותו.

h. לציין האם השתמשתם במהלך העבודה בקוד פתוח לפרט את השירות, כתובת, היכן השתמשתם, כיצד השתמשתם.

i. כמו כן, עליכם להוסיף כל מידע נוסף שלדעתכם חשוב להבנת התכנית ע"י הבוחן.

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 מסמכים 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. הדפיסו את רשימת ה terms של המסמך שמספרו FBIS3-3366 ממוינת לפי הא"ב (לפני stemming) עם תדירות של כל מילה במסמך.

g. הציגו את גודל ה- Posting – נפח האחסון הנדרש עבור קבצי ה- Posting (ב-KB) עבור stemming וללא.

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

בהצלחה!