

## פרויקט גמר – מבוא לתוכנות

בפרויקט זה נרצה לנתח מאגר מידע של התלמידים הבלתי. נתחל עם קובץ אחד `bavli.txt` המכיל את כל הטקסט/global, ונמיין אותו לפי החיפושים שנרצה לבצע בתוך המידע.

### הדגשים לפרויקט:

1. השתמש בקלואסים. הקוד חייב להיות מחולק לקלואסים. כאשר כל פונקציה נמצאת בתוך הקלאס המתאים לה.
2. השתמש בפונקציות. אין לכתוב קוד ארוכים מידי, בכל מצב בו ניתן להוציא קטע קוד לתוך פונקציה\מתודה בצע זאת.
- 3.

### שלב 0:

פתח את קובץ `bavli.txt` עיין בתוכו ולמד את המבנה

### שלב 0.5: כתיבת טסטים

שפות תכנות מאפשרות לנו **לבדיקת הקוד ע"י שימוש במילה שמורה `assert`**  
דוגמא:

```
assert "hello" != "world"  
assert "hello" == "hello"  
assert 1 != 2  
assert 1 == 1  
assert true
```

בצעו קריאות לMETHODES שאתם כותבים וקרוו לassert מול התוצאה הצפוייה  
לדוגמא

```
assert getDaf("מאיימת קוראין").startsWith("א", "ב", "ברכות")
```

## שלב 1.

הדגמה לחיפוש בתוך מידע לא ממוין.

לפני שנמיין את המידע שלנו ונציג אותו באופן יעיל, ננסה לבצע חיפוש בתוך הטקסט הלא ממויין  
הוסף לתוכנית את הפיצ'רים הבאים:

1. קיבל קקלט מהמשתמש מסכתה ודף וחפש אותה בתוך הקובץ לאחר מכן הדפס אותה  
למסך.

יש לנו מצלבים בהם שם המסכת המזון אינו שם תקין

יש לנו מצלבים בהם הדף שהוזן אינו תקין

2. חפש ציטוט מתוך התלמיד  
במידה והציטוט לא נמצא, יש להדפיס הודעה מתאימה

## שלב 2.

מיזן הטקסט לפי דפים

כפי שניתן להבחן, חיפוש בתוך מאגר מידע גדול עם טקסטים לא ממויינים אין ייעיל ודורש זמן  
רב.

בשלב זה נרצה לחלק את תוכן הקובץ הגדול לקבצים המכילים כל עמוד בנפרד.

ניתן לשמר כל קובץ לפי שם המסכת, מספר הדף והעמוד

לדוגמא:

ברכות דף כ"ז עמוד א ישמר בנתיב pages/brachot/26/alef.txt

#### **הוסף לתוכנית את הפיצ'רים הבאים:**

1. צור קלאס המטפל בקריהה של תוכן הקובץ הגדול וכתיבתתו לתוך קבצים קטנים לפי השם
2. קיבל כקלט מהמשתמש מסכתה דף וחפש אותה בתוך הקובץ לאחר מכן הדפס אותה למסך.

יש לנוהל מצבים בהם שם המסכת המזון אינו שם תקין  
יש לנוהל מצבים בהם הדף שהוזן אינו תקין

### **שלב 3.**

#### **טעינת רשיונות ממוחנות של שמורות המסכתות והדפים לתוך הזיכרון**

בשלב הנוכחי יש לנו דרך לבדוק האם מסכתה כל שהיא קיימת בתוך המערכת ע"י גישה למסכת קבצים. הבעיה עם כך היא שהגישה למערכת קבצים היא איטית ובמידה ומתבצעת לעיתים קרובות, יכולה לגזול משאבים מהמערכת.

במקום זאת, נרצה לטען את כל המידע הזה אל תוך הזיכרון בזמן האתחול של התוכנית, ולהחזיק את המידע בזיכרון לאורך חי' התוכנית.

לפנינו שוניגש לבדוק בדיסק אם קובץ כל שהוא קיים, ונבדוק זאת קודם קודם בתוך המבנה נתונים בזיכרון.

#### **הוסף לתוכנית את הפיצ'רים הבאים:**

1. צור מערך ממוחן המחזיק את רשיימת כל שמורות הקיימות במערכת.
2. עבור כל מסכתה החזק מערך ממוחן של דפי הקיימים בהאותה מסכת.
3. כתוב פונקציה המקבלת שם מסכת ודף והמבצע חיפוש ביןאי על רשיימת שמורות על רשיימת הדפים.
4. קיבל כקלט מהמשתמש מסכתה דף וחפש אותה בתוך הקובץ לאחר מכן הדפס אותה למסך.

יש לנוהל מצבים בהם שם המסכת המזון אינו שם תקין  
יש לנוהל מצבים בהם הדף שהוזן אינו תקין

השתמש בקלאס כדי לייצג מסכתה ודפים של מסכתה ושמור את המידע בתוך ליטט.

### **שלב 4.**

#### **יצירת אינדקס עבור ביטויים**

עד לשלב זה יצרנו אינדקסים לחיפוש מסכת ודף  
כעת נרצה לאפשר חיפוש של ביטויים בתוך הדפים

#### 4.1 מציאת כל המשניות של מסכת מבקשת

אנו נרצה לקבל את כל המשניות של מסכת מבקשת. כדי לבצע זאת נצטרך לעبور על כל הדפים של המסכת לקרוא את המידע ולהוציא החוצה את החלקים שהם משניות.

תהליך זה הוא יקר יחסית בהתחשב בעובדה שרוב הדפים אין בהם משניות  
במקומות זה, נרצה להוסיף לכל עמוד מידע האם קיים בתוך מסנה או לא וכן מהקבצים רק את הקבצים בהם יש משניות

בנוסף, נוסיף ערך המציין לאיזה פרק העמוד זהה משוייך

**הוסף לתוכנית את הפיצ'רים הבאים:**

1. סורק את כל הדפים ועדכן את האינדקס עם פרטים לגבי העמודים בהם ישנו משניות  
והפרק עליהם העמוד שייר
2. קיבל מהמשתמש שם של מסכת כל שהיא ופרק והדפס למסך את תוכן כל הגמרא של אותו פרק.
3. קיבל מהמשתמש שם מסכת ופרק והדפס את תוכן המשניות של אותו הפרק.

בונוס. ניתן לציין בתוך האינדקס את מקום תחילת המשנה וסופה וכן לבצע קריאה ישר מהדיסק  
של תוכן המשנה.

### שלב 5. יצירת cache

קריאה מיידית מתוך דיסק איטית מאוד בגין קריאה מהזיכרון.  
כדי לחסוך גישה לדיסק נהוג לשמר על שכבת cache בתוך הזיכרון עבור המידע שנקרא הכי הרבה.

ישנם אלגוריתמים שונים כיצד לנוהל cache, אחד מהם נקרא (LRU) Last Recently Used. נהרגה לשכבת cache בשבייל אחסון דפי גمرا פופולריים.

נרצה להחזיק מערכ דוח מדדי המחזיק טקסט של דפים הפופולריים ביותר בכל מסכת ולנהל את את הדפים/cache לפי הבקשות לצפיה בדף.