

α ה א ו נ י ב ר ס י ט ה ה פ ת ו ח ה

20563

סדנה בבסיסי נתונים

חוברת הקורס קיץ 2025ג

הילה כליף

יולי 2025 - סמסטר קיץ – תשפ"ה

פנימי – לא להפצה.

© כל הזכויות שמורות לאוניברסיטה הפתוחה.

תוכן העניינים

א	אל הסטודנטים
ב	1. לוח זמנים ופעילויות
ג	2. פרויקטים להגשה
ג	3. התנאים לקבלת נקודות זכות בסדנה
ד	4. פרויקט

אל הסטודנטים,

עם הצטרפותכם לסדנה בבסיסי נתונים, אני מאחלת לכם הצלחה רבה, ומקווה שתמצאו בה עניין ותועלת. החוברת שלפניכם כוללת את לוח הזמנים של הסדנה, תנאים לקבלת נקודות זכות והגשת חלקי הפרויקט.

לקורס קיים אתר באינטרנט בו תמצאו חומרי למידה נוספים.
בנוסף, האתר מהווה עבורכם ערוץ תקשורת עם צוות ההוראה ועם סטודנטים אחרים בקורס.
פרטים על למידה מתוקשבת ואתר הקורס, תמצאו באתר שה"ם בכתובת:

<http://telem.openu.ac.il>

מידע על שירותי ספרייה ומקורות מידע שהאוניברסיטה מעמידה לרשותכם, תמצאו באתר הספרייה
באינטרנט www.openu.ac.il/Library.

ניתן לפנות אלי בדואר אלקטרוני: hilak@openu.ac.il, כמו כן אפשר לפנות בטלפון בימי ד' בין השעות 20:30-21:30, בטלפון 053-7307730. במידת הצורך ניתן לתאם פגישה.

בברכה,
הילה כליף
מרכזת הסדנה.

1. לוח זמנים ופעילויות (20563 / 2025)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח ממ"ן (למנחה)
1	13.07.2025-17.07.2025	פרק 5		
2	20.07.2025-24.07.2025	פרק 10		
3	27.07.2025-30.07.2025	פרק 11		
4	03.08.2025-07.08.2025	תכנון הפרויקט		
5	10.08.2025-14.08.2025	פרק 13		
6	17.08.2025-21.08.2025	פרק 12		
7	24.08.2025-28.08.2025.	פרק 24 פרק 9		
8	31.08.2025-04.09.2025	פרק 20		
9	07.09.2025-11.09.2025	מחסון וכריית מידע		

* התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".

2. פרויקט להגשה

בהמשך מובאת הגדרת הפרויקט, שהוא עיקר עבודת הסדנה. יש להגישו בשני חלקים, לפי ההנחיות בהסבר שלהלן.

יש לפנות אלי בדוא"ל לתיאום נושא להרצאה עד סוף השבוע הראשון של הסמסטר!

3. התנאים לקבלת נקודות זכות בסדנה

- א. הרצאה בכיתה.
- ב. הגשת הפרויקט בציון 60 לפחות.
- ג. ציון בקורס 60 לפחות.

הציון הסופי בקורס מורכב מציון הפרויקט בשיעור 70%, ומציון ההרצאה בשיעור 30%.

4. פרויקט - מועד הגשה: 1.1.2026

הגדרת הפרויקט

בפרויקט זה עליכם להקים מערכת שתיתן שירותי קונקורדנציה ואחזור טקסט. הנתונים הגולמיים יהיו קבצי טקסט (מסמכים). הנתונים בטבלאות ייטענו מתוך קובצי הטקסט, עם הכנה מתאימה, וייתוספו להם נתונים נוספים, על פי תבנית בסיס הנתונים (הסכימה) שתכנונו.

המידע הגולמי - הטקסטים

קובצי הטקסט יכולים להיות קבצים כלשהם מסוג *.txt. סביר להניח שיהיה לכם נוח יותר לעבוד עם טקסטים באנגלית, אך השפה היא לבחירתכם. (שילוב שפות שונות עלול לסבך את העבודה). אפשר למצוא טקסטים שיתאימו למטרה זו באינטרנט (ראו הפניות לדוגמה בהמשך).

בסיס הנתונים צריך לאפשר טיפול בטקסטים מכמה קבצים יחד. יש לשמור גם מידע מובנה על הטקסטים השונים, כגון שם הקובץ ומיקומו, וכן נתונים נוספים שהם רלבנטיים לסוג הטקסטים שבחרתם, כגון תאריך הכתיבה, שם המחבר, המקור (אתר האינטרנט או הוצאה לאור).

דוגמאות לטקסטים מסוגים ספציפיים (בין השאר, בהשראת פרויקטים מסמסטרים קודמים):

- **מילות שירים ופזמונים.** במקרה זה הנתונים הנוספים יכולים להיות שם מחבר התמליל, שם המלחין (או כמה מלחינים, אם יש כמה הלחנות), שמות מבצעים וכו'.
- **כתבות בעיתונים.** במקרה זה הנתונים הנוספים יהיו שם העיתון, התאריך, עמוד, שם המחבר (או המחברים), כותרת הכתבה ו/או נושא הכתבה.
- **קוד של תכניות מחשב.** במקרה זה הנתונים הנוספים יהיו נתונים כגון תאריך עדכון, שם התכנית, מספר גרסה, שפה (אם מדובר בתכניות בשפות שונות) וכו'.
- **פסקי דין.** שמות המתדיינים, תאריך, בית המשפט, השופט וכו'.
- **סיכומי מחלה ברשומות רפואיות.** שם החולה ופרטים אישיים נוספים, תאריך, שם הרופא שכתב את הסיכום וכו'.

אין לאחסן את הטקסטים בשלמותם בבסיס הנתונים, אך יש לאחסן מידע על כל אחד מהם, עם זיהוי ייחודי.

הפונקציות השונות בממשק המתמשש:

- טעינת מסמכים (קובצי *.txt).
- הכנסת נתונים מובנים אחרים, בהתאם לסוג המסמכים. (ראו הסבר לעיל)
- אפשרות לשליפת מסמכים לפי הנתונים המובנים, וכן לפי מילים בטקסט.
- הצגת כל המילים בטקסט - תוך אפשרות להציג רק את המילים בקבצים מסוימים, או בכל בסיס הנתונים.

- כאשר בוחרים מילה מהרשימה לעיל - הצגת ההקשר שלה בתוך הטקסט. ההקשר צריך להתבטא בקטע טקסט הכולל שורות לפני ואחרי המילה. עדיף להציג בחלון, כך שהמשתמש יכול "להסיע" קדימה ואחורה, או, אם הצגת ההקשר מוגבלת לקטע קצר - לתת אפשרות קישור לטקסט המלא) אם המילה מופיעה כמה פעמים - יופיעו ההקשרים השונים שלה בזה אחר זה, כך שאפשר לדפדף ביניהם, או לגלול את התצוגה ביניהם.
- הצגת כל המילים כאינדקס: לכל מילה הצגת המיקום שלה. לפחות שני סוגים שונים של הגדרת מיקום. למשל: שורות ועמודים, וכן משפטים ופסקאות. אם בטקסט מוגדר מבנה - אחד המיקומים יהיה על פי המבנה. למשל, אם יש פרקים - נגדיר מיקום על-ידי משפט, פסקה ופרק. אם מדובר בשירים עם בתים - נגדיר שורה בבית. או מדובר בקבצים של תוכניות מחשב - אפשר למפות לפי מודולים, מחלקות, פונקציות וכו' - לפי המבנה.
- אפשרות לאתר מילה לפי המיקום שלה. (כגון: מצא את המלה המופיעה בשורה 20 בעמ' 12 וכו')
- מתן אפשרות למשתמש להגדיר קבוצות של מילים, בעלות משמעות מיוחדת ולהציג את המופעים שלהן בטקסט. יכולות להיות קבוצות שונות, בעלות שמות שונים, ושיוך המילים אליהן ייעשה על ידי הזנה מפורשת, או על ידי סימון מילים מתוך בסיס הנתונים כשייכות לקבוצה. המשתמש יוכל להגדיר קבוצה, ואז להכניס לתוכה מילים באחת הדרכים (או בשתייהן). למשל: קבוצת המילים שהן שמות של חיות; קבוצת מילים שהן פעלים וכו'. המשתמש יוכל להגדיר בכל פעם קבוצה חדשה, לתת לה שם, ולהוסיף לה מילים בכל עת.
- מתן אפשרות למשתמש לאחסן ביטויים לשוניים (על פי הגדרת המשתמש), עם הופעותיהם בטקסט. "ביטוי לשוני" הוא בעצם רצף סדור של מילים. המשתמש יוכל להגדיר ביטוי ואז לחפש אם הוא נמצא בטקסט, או לסמן ביטוי בתוך הטקסט ולבקש את הופעותיו הנוספות.
- עבור קבוצת מילים שהוגדרה במסגרת הגדרת קבוצות - אפשרות להציג אינדקס רק למילים באותה קבוצה. למעשה, שירות זה הוא הצגה של תת-קבוצה מתוך רשימת המילים הכללית. הרעיון הוא לאפשר למשתמש (מבחינת הממשק המוצג לו) לייצר אינדקס כזה כדו"ח להדפסה, או כקובץ במערכת, עם שם מתאים. למשל, אינדקס של שמות מקומות.
- נתונים סטטיסטיים: מספר תווים במילה, במשפט, בעמוד, בפרק, בקובץ; מספר מילים במשפט, בפרק וכו'; רשימות שכיחות למילים.

יישום אחד הנושאים שנסקרו בהרצאות

במסגרת הפרויקט תצטרכו להראות כיצד יישמתם את אחד הנושאים שנסקרו בהרצאות. יש כמה אפשרויות עיקריות:

- כוונון - התייחסות לביצועי המערכת. כדי ליישם היבטים של כוונון יש להציג את עבודת המערכת עם מאגר נתונים גדול יחסית, ולהציג את שיקולי הביצועים שלקחתם בחשבון. במידת האפשר, יש לתת הצגה השוואתית של ביצועי המערכת, עם ובלי ההיבטים שיישמתם.
- XML - יכול להיות שימושי בפרויקט בכמה אופנים. למשל, לצורכי גיבוי, לייצא את הנתונים ל-XML ואז לבנות את בסיס הנתונים מחדש מתוך יבוא של ה-XML שנבנה. או לאפשר העברה של המערכת לתוכנת ניהול בסיסי נתונים אחרת, ושימוש בה שם.

- כריית מידע: אפשר להפעיל פעולות של כריית מידע על הטקסטים בעזרת כלים קיימים, או לממש באופן ישיר את אחד האלגוריתמים הבסיסיים - למשל אלגוריתם אפריורי למציאת חוקי הקשר. (למשל: זוגות של מלים המופיעות יחד לעתים קרובות בתוך משפט אחד, או בתוך פסקה אחת וכו')
- פונקציונליות מיוחדת נוספת בתחום הטקסטואלי: למשל, ניתוח דקדוקי של המילים כך שאפשר לחפש מילים לפי צורתן השורשית ולקבל גם הטיות שונות שלהן. נושא זה אינו יישום של אחד מנושאי ההרצאות, אך מכיוון שעשוי להיות בו עניין ליישום עצמו, גם זו אפשרות.

שלבי העבודה

א. תכנון

בשלב ראשון יש לנתח את המערכת ולהגדיר את תבנית בסיס הנתונים (הסכימה) באופן שיאפשר את מתן השירותים הנ"ל. יש להגיש בנפרד מסמך שיבטא את הניתוח הזה. המסמך צריך להכיל תיאור מילולי של המערכת שאתם מתכננים, דיאגרמת ישויות קשרים לתיאור הסכימה של בסיס הנתונים, וכן הגדרה של הטבלאות והאילוצים החלים עליהן ב-SQL. כמו כן יש לכלול בתכנון תיאור של מרכיבי ממשק המשתמש, רצוי עם תיאור גרפי של חלקם לפחות, וכן דוגמאות לשאילתות SQL שהמערכת שלכם תעשה בהן שימוש.

את התכנון יש לשלוח אליי למייל hilak@openu.ac.il

המשך הפרויקט יבוצע לאחר אישור התכנון.

ב. מימוש

השלב השני יכול את מימוש המערכת כולה. בסיס הנתונים עצמו יוגדר במערכת Oracle (או במערכת אחרת - בתיאום עם מרכזת ההוראה). יהיה עליכם לכתוב גם תוכניות לסריקת קבצי הטקסט, פירוקם למילים (או כל עיבוד הדרוש על פי תכנונכם), טעינת הנתונים מתוכם ואכלוס הטבלאות על פי התבנית שהוגדרה. את התוכניות תכתבו בשפה כלשהי הנוחה לכם, או בכל כלי אחר המאפשר קישור לבסיס הנתונים. יהיה עליכם לממש במערכת את הפונקציונליות של השירותים שהוגדרו לעיל, ולבנות ממשקי משתמש מתאימים לביצוע שירותים אלה.

העבודה תיעשה בזוגות בלבד.

הגשת הפרויקט מורכבת משני חלקים :

1. הגשת מסמך כתוב שמפרט את הסכמה, תהליך העבודה, הכלים שנבחרו, הנושא הנוסף ותיאור קצר של הבעיות בהן נתקלתם ופתרון.

2. **הצגת הפרויקט בפני צוות הקורס.** לצורך כך, ייקבע מפגש הדגמה בין הסטודנטים המגישים את הפרויקט למרכז ההוראה.

הצגת הפרויקט צריכה להיות מלווה במצגת המתארת את :

- הסכמה
- תהליך העבודה
- הכלים שנבחרו
- הנושא הנוסף
- הבעיות שנתקלתם בהן ופתרון

ד. ניקוד

10% - ההצגה עצמה

15% - הנושא הנוסף

15% - כל תפקודי המערכת (קבוצות, יחסים, הקשר וכו')

15% - ממשק משתמש

45% - סכימה, התרשמות כללית

מערכות דומות להתרשמות והשראה

- תוכנה מסחרית הנותנת שירותי קונקורדנציה :

<http://www.concordancesoftware.co.uk>

- תוכנה חופשית עם שירותים דומים :

<http://www.textworld.com/index.html>

אתרים עם טקסטים :

- Electronic text center :

<http://etext.lib.virginia.edu>

- Internet Public Library :

<http://www.ipl.org/>