מכללת הדסה, החוג למדעי המחשב

אנליזה של ביג דאטה / חננאל פרל

סמסטר א', תשפ"ה

תרגיל גדול אחרון / 14.01.2025 - עדכון 16.01.2025 - 21:55

תאריך הגשה:

הגשת סופית של כל החלקים: יום שני 3.02.2025 בשעה 23:00

<u>הצגה חובה פיזית פרונטלית בכיתה</u>: בשיעור האחרון יום רביעי 5.02.2025 בשעה 17:00.
הצגת התרגיל היא במשך כ-8 דקות (4 דקות הצגה לכל אחד מבני הזוג).
אי הצגה בכיתה תגרור הורדה בציון של רכיב זה.

כל מי שלא יכול להגיע פיזית ולא יכול להציג, צריך לקבל אישור כתוב.
יש לבקש בכתב <u>עד</u> יום שני 3.02.2025 בשעה 18:00. לאחר מועד זה לא יתקבלו בקשות.

מטרות התרגיל:

תרגיל גדול אחרון - התרגיל מדמה מערכת אמיתית שבה יש דאטה גדול, מהנדס וחוקר הנתונים מוציאים ממנו תובנות מעניינות שנשמרות כדאטה קטן, ואז מוצגות למשתמש הקצה בדאשבורד.

הוראות ודרישות:

כל זוג יצטרך להגיש מסמך קצר המתאר ומסביר את התרגיל המסכם, יש לציין ולתאר בצורה ברורה את האחריות והעבודה של כל אחד מהסטודנטים במהלך התרגיל.

חיפוש באינטרנט ובחירת מאגר נתונים גדול – עשרות מיליוני שורות, הרבה עמודות, ולפחות 1 גיגה (GB 1) גודל של הקבצים פתוחים (לא דחוסים).

הכי נח זה מאגר שיופיע כקובץ CSV, יתכן ויהיו כמה קבצי CSV ביחד שיגיעו לגודל הרצוי. (בכל מקרה לא קבצי JSON).

כל זוג צריך לבחור מאגר שונה.

יש לקבל אישור שלי על המאגר לפני תחילת עבודה עליו.

הגשת בקשה לעבודה על מאגר תתבצע כאן:

https://docs.google.com/forms/d/e/1FAIpQLSdIUtWnvO4GCv4Y1PtcDmEeE07ggePDnM10lU2JHyX6Yap69w/viewform?usp=sf_link

יש לעקוב בלינק התגובות (כל תלמיד רואה את כל הבקשות בכיתה) ולראות האם המאגר הזה כבר נבחר על ידי מישהו אחר, ולראות האם המאגר אושר עבורך או לא (עמודות H I) ולענות לשאלות לפי הצורך (בגיליון עצמו אין אפשרות עריכה יש רק אפשרות תגובה):

https://docs.google.com/spreadsheets/d/1nviWO8cwRnDq_u7bgB5m4Nd8V49NPpvJ8By3jNfrbxc/edit?usp=sharing

המאגר הנבחר הזה ישמש בתרגיל והוא יהיה כמעין ביג דאטה Warehouse

יש להגיש (1) לינק למקור הנתונים, תאור של המאגר מבחינה טכנית

(2) יש לציין מה גודלו, כמה שורות? כמה עמודות? וכמה סך נפח המידע?

(3) וגם לתאר את התוכן של המאגר מה יש במאגר? מה התוכן הכללי שלו..? מאיפה הוא הגיע וכו.. ולכלול הסבר מפורט על עמודות חשובות.

נוריד (אפשר ידנית) מהאינטרנט את הקבצים הגדולים.

(4) אין להגיש את הקבצים הגדולים עצמם שהורדתם מהאינטרנט. עבור כל קובץ CSV (או משהו אחר כאמור) שהורדתם מהאינטרנט, יש לייצר קובץ מקביל עם רק 500 שורות (עמודות להשאיר זהות), ואותו כן להגיש אותו. [זה לצורך דוגמא לבודקת שתראה איך נראה הקובץ]

באמצעות DUCKDB נכניס אותם לתוך קובץ דאטהבייס - db_file.duckdb
https://duckdb.org/docs/guides/file_formats/overview

(5) יש להגיש תאור מילולי ואת הפקודות שהרצת כדי להעלות את הדאטה הגדול לתוך הביג דאטה DUCKDB.

שלב הבא יהיה להבין את הדאטה ולבחון אותו, לראות מה יש בו ומה אין בו. מה המאפיינים של הדאטה ומה אני יכול ללמוד מהם, אלו תובנות אני רואה, אלו סטטיסטיקות אפשר להראות, איזה עמודות מכילות מידע חסר, וכדומה..

נשאל שאלות על הדאטה (4-6 שאלות)

נכתוב שאילתות SQL שעונות על השאלות

נחשוב מה הסיפור שנרצה לספר? איך נספר אותו?

נציג ויזואליזציות כחלק מהסיפור..

את השאילתות על הביג דאטה נכתוב עם SQL ע"י כל סט היכולות המוכרות, פונקציות אגרגציה, חלון, OLAP, וכו יש לוודא שמופיע בשאילתות גם קבוצות: GROUP BY | HAVING וגם חלון: OVER | PARTITION.

העבודה תהיה עם DuckDB התקנה והרצה לוקלית, לינק: https://duckdb.org דרך פייתון כמובן.

שימו לב יש להשתמש במגוון היכולת של SQL מעל DUCKDB (שהוא כזכור יותר עשיר מאשר SQLITE) למשל CUBE, PIVOT וכו..

את התוצאות של הריצה של שאילתות SQL על הביג דאטה נשמור בטבלאות קטנות יותר גם בתוך ה DUCKDB.

חוץ מזה, יש לייצר עבור כל טבלה גדולה, טבלה "קטנה" עם בערך 500 שורות מאפיינות ומייצגות את הנתונים מתוך כל טבלה גדולה. https://duckdb.org/docs/sql/samples.html

(בעצם, בפועל, השאילתות שעונות על השאלות, ומכינות הדאטה לויזואליזציות, מחזירות דאטה קטן במגוון שיטות: אגרגציה, פילטור, Samples וכו' לפי מה שלמדנו. יש לבחור כמה מהשיטות ולא רק שיטה אחת.)

(6) יש להגיש את השאלות ששאלת על הדאטה (4-6 שאלות) ומה אני יכול ללמוד מהם, אלו תובנות אני רואה.

(7) יש להגיש את השאילתות SQL שעונות על השאלות, כולל תאור מילולי (למעשה שאילתות אלו מייצרות את הטבלאות "הקטנות". יש לתת הסברים אלו שיטות "להקטנת" הדאטה נבחרו כל פעם ולמה.

(8) יש להגיש את השאילתות שבוחרת שורות לדוגמא מכל טבלה.

לבסוף נעתיק את כל הטבלאות "הקטנות" שיצרנו ונשמור בתוך SQLITE המוכר, לינק: https://www.sqlite.org , גם כאן דרך פייתון כמובן.

הדאטה בייס הקטן יכול להכיל כמה טבלאות.

העברת המידע תתבצע על ידי DUCKDB באמצעות - SQLITE extension
https://duckdb.org/docs/extensions/sqlite.html

(9) יש להגיש תאור מילולי ואת הפקודות שהרצת כדי להעלות את הדאטה הקטן לתוך SQLITE.

קוד פייתון

(10) יש להגיש את הקוד פייתון שמריץ את כל התהליך עד כה

(11א) יש להגיש את הקובץ דאטהבייס "הקטן" בעצמו - את SQLITE.

(11ב) יש להגיש רשימת כל הטבלאות שנמצאות בדאטה בייס הקטן (ה SQLITE) כמה שורת וכמה עמודות בכל טבלה, והסבר קצר מה יש בטבלה ולאיזה צורך.

נייצר לבסוף **דאשבורד** על ידי קוד פייתון עם StreamLit לינק: https://streamlit.io . נא להשתמש ברכיב המובנים https://docs.streamlit.io/develop/api-reference ולא בצד שלישי "Third-party components".

הדאשבורד יקרא את כל הנתונים שלו, רק מתוך הטבלאות הקטנות, ז"א רק מתוך SQLITE. (הדגשה: הדאשבורד לא פונה ל DUCKDB)

להראות במסכים של הדאשבורד את הבערך 500 שורות מאפיינות לדוגמא ששמרתם ב SQLITE.
יש להראות את זה כטבלה עם צבעים. להשתמש ב st.dataframe ואת הצבעים על ידי df.style.

לכתוב במסכים של הדאשבורד את השאלות את הסיפור ואת התובנות

וגם להוסיף 5 גרפים / ויזואליזציות:

2 הגרפים חובה שיהיו של https://matplotlib.org - Matplotlib/

1-2 גרפים אפשר שיהיו https://seaborn.pydata.org - Seaborn/

2 גרפים חובה שיהיו אינטראקטיביים, ז"א היוזר יוכל לבחור משהו במסך (למשל מתוך תפריט, סליידר, וכו) ואז לקבל גרף שמותאם לבחירה.

אפשר שחלק מהויזואליזציות האלו יהיו מסוגים נוספים: טבלה עם צבעים, word cloud ענן מילים, מפות גיאוגרפיות, ודברים נוספים לאו דווקא גרפים במובן הפשוט של המילה. ז"א אפשר דברים יותר יצירתיים לאו דווקא מתוך הסיפריה של Matplotlib.

כזכור התוכן של כל הגרפים, גם האינטרקטיביים, ילקח אך ורק מהדאטה בייס הקטן SQLITE.

יש לבחור גרפים מסוג דו מימד 2D בלבד (ולא תלת מימד)

(12) יש להגיש ולכתוב מה הסיפור שנרצה לספר? איך נספר אותו?

(13) יש להגיש תאור סכמתי איך יראה הדאשבורד שלנו

(14) יש להגיש הסברים אלו וויזואליזציות נבחרו, למה ומה רואים בהן..

(15) יש להגיש את הקוד פייתון של האפליקצייה (אפשרי כמה קבצים) שמצייר ומציג הדאשבורד עם הטבלאות לדוגמא ואת הויזואליזציות (לא לשכוח מקרא). הדאשבורד יכלול גם הסברים על הסיפור מה רואים וכו.

(16) יש להגיש כל קובץ נוסף שצריך כדי שנוכל להריץ את הדאשבורד לוקלי אצלנו. בסופו של דבר הבודקת מורידה את הקבצים שהגשתם ומריצה את הדאשבורד אצלה, זה חלק מהבדיקה.

(17) יש להגיש קובץ requirements.txt המכיל את כל חבילת הפייתון הנצרכות להרצה.

(18) יש לכתוב הסברים איך להריץ.

GRPAH DB - אילו הייתם צריכים לשמור ולמדל את הנתונים שהורדתם בחלק א בתור GRPAH DB.
איך הם היו מסודרים שם?

(19) יש לצייר סכמה (כמו מה שהפקודה הזו מציירת: CALL db.schema.visualization())

הסכמה המצוירת תכלול את ה NODES וגם את ה RELATIONSHIPS.
יש צורך גם לציין מה יהיו ה PROPERTIES של כל אחד מהם.

(20) יש צורך גם להסביר ולפרט באמצעות טקסט קצר את הסכמה ומה שנמצא בה.
אין חובה למדל את כל העמודות בקובץ, מספיק להראות כ-8 עמודות..

(21) - (24) אולי אולי יהיו תוספות הקשורות ל **Data enrichment** - נרצה להוסיף מידע על הדאטה הקיים ע"י SCRAPER. ואולי יהיו תוספות הקשורות ל **SPARK**.. כרגע להשאיר ריק.. לא יהיו כאן תוספות

בנוסף יש להגיש:

(25) צילומי מסך של כל הדפים של הדאשבורד.

(26) קטע קצר המתאר ומסביר את התרגיל המסכם, מבחינת ארכיטקטורה וקוד, כולל ציור סכמתי של המערכת.
אפשר לצייר בעזרת https://www.drawio.com

(27) תיאור בצורה ברורה את האחריות והעבודה של כל אחד מהסטודנטים במהלך התרגיל.

(28) קובץ **README** – שמות המגישים, ורשימת כל הקבצים שהוגשו (כל הקבצים שבתוך הZIP), והסבר קצר מה יש בכל קובץ.

כל ההסברים והתאורים והתמונות צריכים להופיע בקובץ וורד (docx) אחד.

ההגשה בסוף תהיה קובץ ZIP עם כל מה שמבוקש למעלה (1)-(28).

דגשים:

- יש לוודא שהדאטה הקטן אכן מכיל מינימום דאטה הנדרש להצגה בדאשבורד. רק שורות נצרכות, רק עמודות נצרכות ושאין כפילויות משמעותיות של דאטה בכמה טבלאות וכדומה..
- יש לייצר את הדאטה הקטן בצורה נקייה על ידי יצירת העתקת הטבלאות לדאטהבייס חדש (ללא מחיקות בדרך). אם כותבים ומוחקים הרבה פעמים הדאטהבייס יהיה הרבה יותר גדול ממה שצריך.
- אם רואים שהביצועים של הדאשבורד על הדאטה הקטן מאד איטיים בטבלה או טבלאות, נא להוסיף אינדקסים ב SQLITE כדי לשפר הביצועים.
- אין להגיש את הקובץ (או קבצים) של הדאטה הגדול שהורדתם מהאינטרנט! גם אין להגיש את הדאטהבייס הגדול של DuckDB (הגשה שלהם תוריד בציון..)
- כל קובץ שמופיע בהגשה חייב להיות מתואר בקובץ README.

הגשה:

יש להגיש במודל קובץ ZIP אחד הכולל הכל.

אם גודל הקובץ ZIP חורג מהמגבלה של 50 MB ולא מצליחים להגיש במודול, אז יש לשלוח אלי אימייל בהקדם..!

כמו בפרויקט בעבודה ובחיים האמיתיים, יתכנו שינויים ותיקונים במהלך התרגיל, אנו עקבו אחר ההודעות!

בהצלחה לכולם ! בהצלחה לכולן !