**מדריך פרויקט סופי בSQL**

**הקדמה**

מדריך זה נועד להסביר, שלב אחר שלב, כיצד לבצע את הפרויקט הסופי בקורס, כך שניתן יהיה לשחזר את התהליך בקלות.

המדריך מתאר את סדר הפעולות המומלץ – החל מהקבצים שיש להריץ לפני תחילת העבודה, דרך יצירת המודלים, איגודם, ועד להרצת בדיקות הביצועים באמצעות **Confusion Matrix**.

המטרה היא לספק תהליך ברור, עם דוגמאות והסברים לכל שלב, כך שכל סטודנט יוכל להבין לא רק *מה* צריך להריץ, אלא גם *למה*.

בפרויקט זה אנו מתנסים בעולם של **מערכות המלצה (Recommendation Systems)** – תחום מרכזי במדעי הנתונים, בו אנו מפתחים מודלים המציעים פריטים (במקרה שלנו – סרטים) בהתבסס על קשרים וחפיפות בין מאפיינים שונים, כגון במאים, שחקנים, וז’אנרים.ֵ

**מה להריץ לפני שמתחילים (וסדר ריצה)**

**הערות כלליות:**

* **ודא/י קודם:** USE imdb\_ijs;
* רצוי להריץ כל קובץ כפי שהוא מופיע בתיקיית ה-Git של הקורס
* יש להריץ 4 קבצים
  1. build\_gs
  2. movies\_recomendations
  3. movies\_recomendations\_agg
  4. build\_restricted\_gs\_db
* סדר חובה:

build\_gs -> movies\_recommendations -> movies\_recommendations -> build restricted\_gs\_db

* אם מריצים מחדש - תנו לסקריפטים להפיל וליצור טבלאות מחדש (ה-DROP כלול בקבצים של המרצה).

1. [**build\_gs**](https://github.com/evidencebp/databases-course/blob/main/recommendations_goldstandard/build_gs.txt)

**מה זה עושה:**

* יוצר את טבלת ה-Ground Truth של הקורס: movies\_recommendations.
* מגדיר מפתחות ראשיים/זרים ואילוצי תקינות (כולל CHECK על טווח ציון 1–10).
* מוסיף דוגמאות ראשוניות (seed) כדי שתהיה תשתית בסיסית גם בלי תרומות הסטודנטים.

**למה צריך:**

* זו “האדמה” שעליה מודדים Precision/Recall. בלי זה, אין למה להשוות.

1. [**movies\_recommendationsֵ**](https://github.com/evidencebp/databases-course/blob/main/recommendations_goldstandard/movies_recommendations.sql)

**מה זה עושה:**

* מוסיף לטבלת movies\_recommendations את כל ההמלצות שהכיתה הגישה (בציון 1–10 + נימוק).

**למה צריך:**

* כדי שה-Ground Truth יהיה אמיתי ומלא – לא רק ה-seed. זה הדאטה האמיתי להשוואה.

1. [**movies\_recommendations\_agg**](https://github.com/evidencebp/databases-course/blob/main/recommendations_goldstandard/movies_recommendations_agg.sql)

**מה זה עושה:**

* מחשב אגרגציה לכל זוג (A,B) מתוך movies\_recommendations, ויוצר movies\_recommendations\_agg עם ממוצע ציון, סטיית תקן, ועוד.

**למה צריך:**

* ה-confusion matrix עובד מול גרסה מאוגדת (זוג אחד ייחודי, לא כפילויות של כמה מציעים).
* זה ה-“class” שלנו בהשוואה: “מה נכון לפי הכיתה”.

1. **ֵ**[**build\_restricted\_gs\_db**](https://github.com/evidencebp/databases-course/blob/main/recommendations_goldstandard/build_restriced_gs_db.txt)

**מה זה עושה:**

* בפועל זו “תת-תצוגה” של ה-IMDB שמכילה רק את מה שבאמת צריך כדי להריץ מודלים במהירות.ֵ
* בונה סכמה מצומצמת של הטבלאות רק עבור הסרטים שרלוונטיים ל-GS (ול-personal ranking אם יש):
  + gs\_movies, gs\_roles, gs\_movies\_directors, gs\_directors, gs\_movies\_genres, gs\_actors.

**למה צריך:**

* מריץ את המודלים מהר יותר, עם פחות רעש ובעיות ביצועים.
* מקטין סיכוי לשגיאות על סרטים שאין להם הקשר ב-GS.

**הערה חשובה:**

* ניתן להריץ עכשיו לצורך שיפור ביצועים אך יהיה צורך בהרצה מחודשת לאחר מטלת הבונוס.

**יצירת המודלים – מה הרעיון**

בונים **שאילתות** (אפשר להשתמש ב- Views) שמייצרות **זוגות סרטים** לפי היגיון שתבחרו. כל מודל = היגיון אחר.

לרעיונות הקש [כאן](https://www.google.com/url?q=https://github.com/evidencebp/databases-course/blob/main/Exams%2520and%2520assignments/Questions/data_integrity_questions.pdf&sa=D&source=docs&ust=1755015780768411&usg=AOvVaw2rtSJUuqGFukKRp20bMg6P)

לרעיונות מתקדמים הקש [כאן](https://github.com/evidencebp/databases-course/blob/main/Exams%20and%20assignments/Questions/Recommendation_ideas.pdf)

(רק כדוגמה, אתם מחליטים את הספים/תנאים):

* **אותו במאי** (למשל, סרטים של אותו במאי בהפרש עד 5 שנים).
* **שחקנים משותפים** (למשל, לפחות N שחקנים משותפים).
* **ז’אנרים משותפים** (למשל, לפחות ז’אנר אחד משותף).

אפשר כמובן לשלב בין תנאים (למשל במאי משותף + ז’אנר משותף), הכול לפי איך שאתם רוצים לכוון את התוצאות.

**איחוד המודלים**

אחרי שיש כמה מודלים, מאחדים את כל הזוגות לרשימה אחת (איחוד פשוט).

אם אותו זוג מופיע בכמה מודלים, אפשר:

* **לשמור דירוג קבוע (fixed)** לכל מודל, ואז לחשב **ממוצע** לזוג שהופיע בכמה מודלים.
* או פשוט בשלב האיחוד לתת דירוג **10** לכל זוג.

חשוב: כדי שהקוד של ה**Confusion Matrix** יזהה “חיובי”, הדירוג הסופי צריך להיות **יותר מ-5**.

לכן, אם אתם נותנים **דירוג קבוע**, דאגו שהוא יהיה **6 ומעלה.**

**הרצת ה Confusion Matrix**

לאחר שיש לכם את טבלת ההמלצות המאוחדת (עם שדה recommendation), פשוט מריצים את קובץ ה-**confusion matrix** שהמרצה סיפק.

מה שצריך לשים לב אליו:

* בקובץ יש חלק שמגדיר **מאיזו טבלה לקרוא את ההמלצות** (בדרך כלל בשורת FROM ... ב-predictor).
* שנו שם כך שיתאים לשם הטבלה/ה-View הסופי שלכם (למשל my\_models\_agg או כל שם שנתתם).
* שאר הקוד של ה-confusion matrix לא דורש שינוי.

אחרי הרצה, תקבלו את מדדי הביצועים (Precision, Recall, וכו’) ותוכלו לנתח את איכות המודלים שלכם.

**פורמט ההגשהֵ**

בסיום העבודה יש להגישֵ שני קבצים:

1. **קובץ SQL** – יכלול את כל שאילתות המודלים שכתבתם, כולל ה-UNION שמאגד אותם לטבלה/תצוגה אחת.
2. **קובץ PDF** – יכלול:
   * הסבר קצר על ההיגיון של כל מודל שבניתם.
   * ניתוח קצר של תוצאות ה-Confusion Matrix (לדוגמה: באיזה מדדים המערכת חזקה, ומה ניתן לשפר).

הערות חשובות:

1. ניתן לעשות את הפרוייקט בזוגות, בנוסף למילואמניקים מותר להצטרף לשלישיה.
2. **כל חברי הקבוצה צריכים להגיש את אותם קבצים**. (ייתכן ציון שונה עקב ציון שונה על הGS)
3. שם הקבצים צריכים לכלול את שם הקבוצה.
4. בקובץ pdf יש לציין את תעודות הזהות של חברי הקבוצה.

**קישור למכווןֵ הבדיקה נמצא** [**כאן**](https://github.com/evidencebp/databases-course/blob/main/recommendations_goldstandard/project_checking_guidlines.pdf)**.**