שם הקורס: ניהול מידע מבוזר (096224)

מגישים: נימרוד סולומון (ת.ז 206574733) ומתן שילוני (ת.ז 208634469)

מספר תרגיל הבית: פרויקט חלק א', חלק 1

תאריך הגשה: 09.06.2022

חלק א' – Extract and Transform

במסמך זה נפרט על אודות הטרנספורמציות שביצענו על קבוצת הרלציות המקורית ועל אודות השיקולים שהנחו אותנו בעבודה על חלק זה.

עקרונות מנחים

1. **רלציות בהן לא שינינו דבר:**
   א. Tickets
   ב. Users

   מבנה הרלציות לעיל היה מסודר דיו לטעמנו, ולכן לא ראינו צורך לבצע עליהן טרנספורמציות.
2. הרלציה הראשית עליה הסתכלנו וממנה נגזרו כל פעולותינו בחלק זה היא queries. זאת מתוך הנחה שהפיצ'רים שלה הם היחידים שמסד הנתונים ניגש אליהם בעת ביצוע שאילתות. לכן, ביצענו טרנספורמציות על פיצ'רים מקבילים ברלציות אחרות, במטרה להקל על נראותן (לצורך המשימות הבאות בפרויקט עבורנו המתכננים) ועל אופן הגישה לפריטי מידע אלו. כמו כן, השארנו פיצ'רים מסוימים בתוך רלציות שביצענו עליהן טרנספורמציות כפי שסופקו גם אם המבנה הסכמתי של שאילתות לא מאפשר גישה אליהן.
3. **סוגי טרנספורמציות שביצענו:**
   א. <u>פיצ'רים המכילים מחרוזת מהצורה ['Family', 'Action']</u>
      1) המרת הסכמה של העמודה מ-string ל-array המכיל strings.
      2) הסרת תווים מיותרים (למשל '...').
      3) לאחר הטרנספורמציות קיבלנו פיצ'ר המכיל מערך מהצורה [Action, Family].
   ב. <u>פיצ'רים המכילים מחרוזת מהצורה [{'id': 35, 'name': 'Romance'}, {'id': 10749, 'name': 'Comedy'}]</u>
      1) המרת הסכמה של העמודה מ-string ל-array המכיל strings.
      2) חילוץ הערכים הרלוונטיים (בהתאם לרלציה queries) וסילוק ערכים שאינם אינפורמטיביים.
      3) לאחר הטרנספורמציות קיבלנו פיצ'ר המכיל מערך מהצורה [Comedy ,Romance].
4. **הנחות:**
   א. מסד נתונים מסוגל לשלוף נתונים מרלציות המכילות פיצ'ר מסוג array of strings.
   ב. אנו מודעים לכך שייתכנו כפילויות בפיצ'רים שהסרנו מהם את הערכים המזהים (מחרוזות id, קיצורי שמות מדינה וכו'). אולם, אנו מניחים כי מידע זה אינו מספק תרומה אינפורמטיבית בנוסף לערכים שהשארנו.

כדי לראות את הרלציות לאחר הטרנספורמציות שפורטו לעיל ניתן להריץ את קובץ הפייתון המקביל שנדרשנו להגיש בחלק זה.