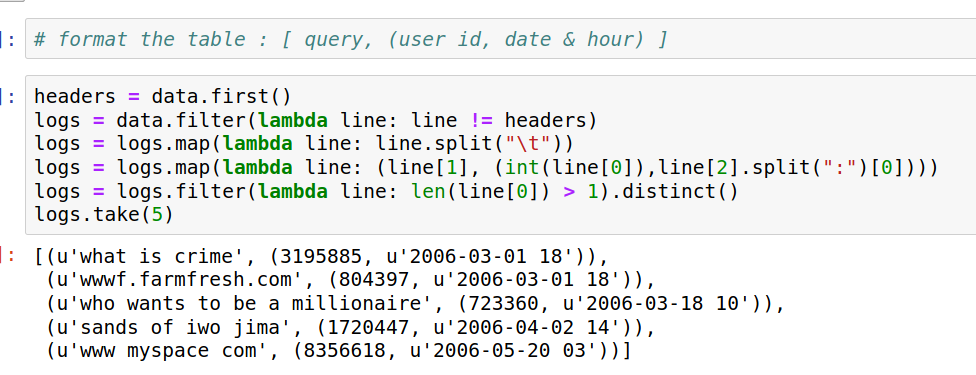
**פתרון ש.ב :אוהד קשת 205926215 ועדן פחימה 20603558**

**אנליזה של ביג דאטה תר' מס' 2:**

בתרגיל זה התבקשנו לממש את האלגוריתם Associative rule learning אשר נלמד בהרצאות בצורה היעילה ביותר, ולממשו על הקובץ טקסט אשר קיבלנו. ובסופו להציג קורלציות מתאימות כנדרש.

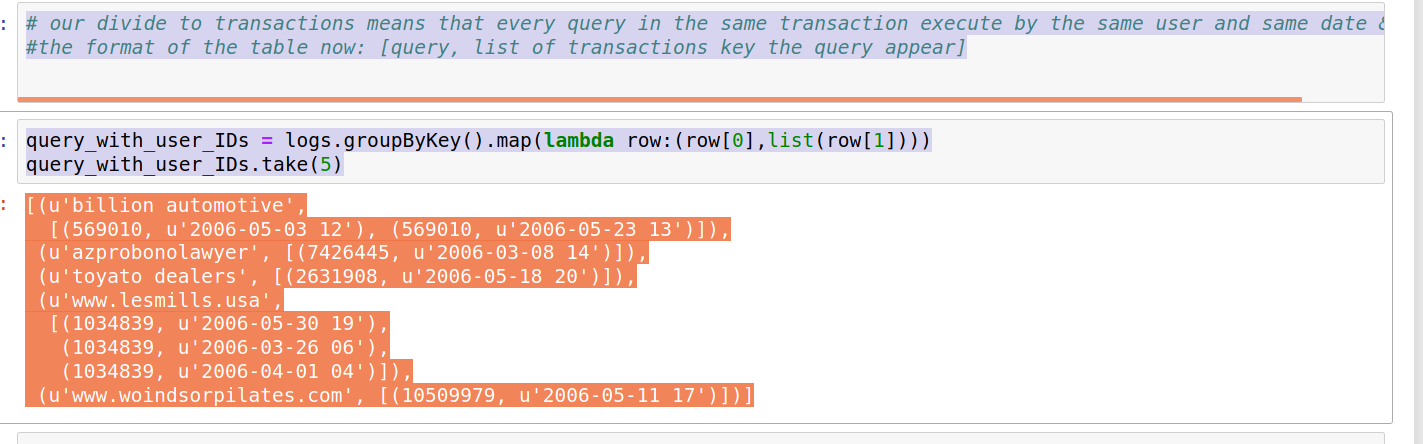
**שלב ראשון:** מסדרים את הקובץ טקסט לפורמט מתאים לעבודה בצורה של key,value כך שהמפתח הוא השאילתה והערכים הם הid של המשתמש התאריך והשעה(בה בוצע השאילת)



**שלב שני:**

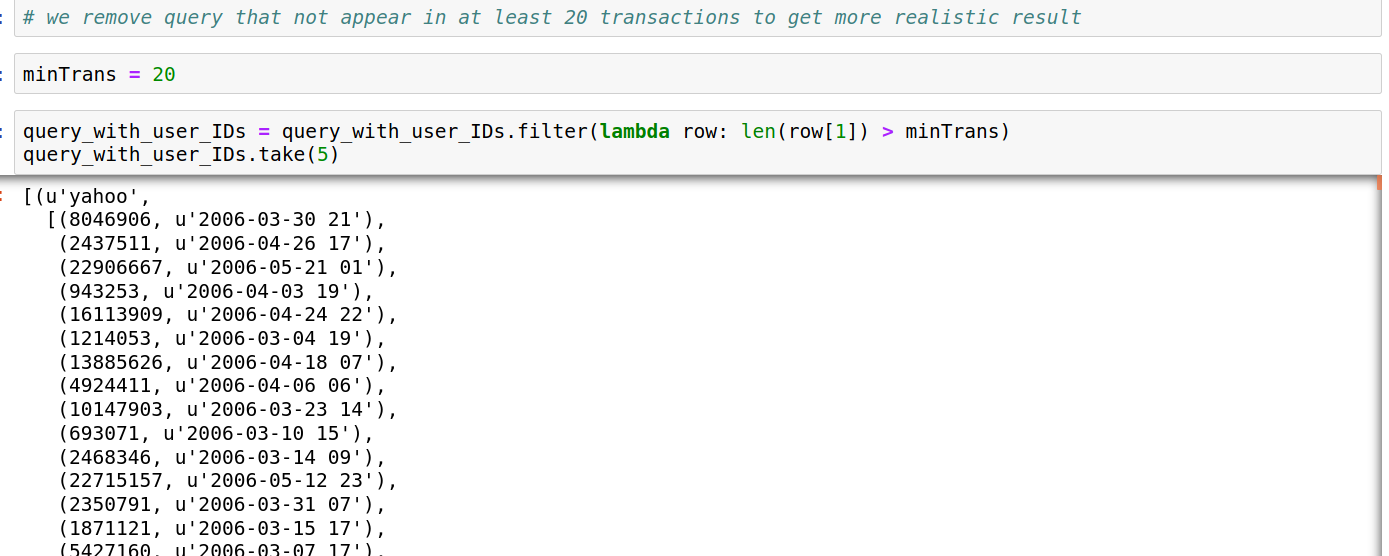
אנו מחלקים את כל השאילתות לטרנזקיות, כך שבכל טרנזקציה כל שאליתה מבוצעת על ידי אותו משתמש באותה תאריך ושעה

כעת נקבץ את השאילתות לקבוצות לפי השאילתה כלומר הפורמט של הטבלה כרגע הוא: שאילתה,ורשימת הטרנזקציות שהשאילתה מופיע בהן



**שלב שלישי:**

הסרנו שאילתות שאינן מופיעות לפחות ב20 טרנזקציות כדי לקבל תוצאות יותר מוחשיות ושניתן לעבוד איתן (במקביל ל minsup במאמר)



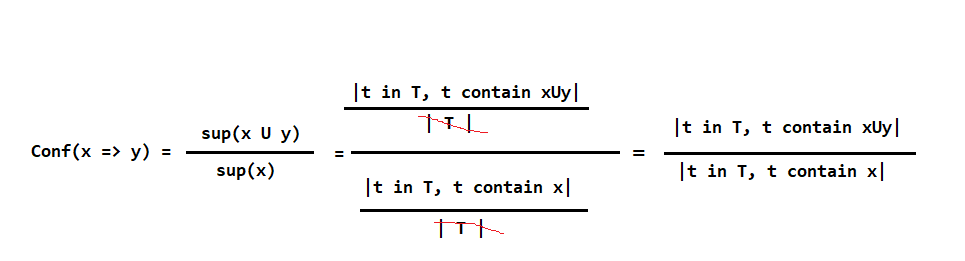
**שלב חמש+שש:**

כעת לאחר שסידרנו את הדאטה שלנו לצורה יותר נוחה לעבודה והסרנו טרנזקציות שאינן ריאליות לחיפוש שלנו.נתפנה כעת לממש את האלגוריתם שלנו.

יצרנו טבלה עם כל צירוף של קומבינציה אפשרית של שאליתות(כך נוכל למצא את הconfidence)

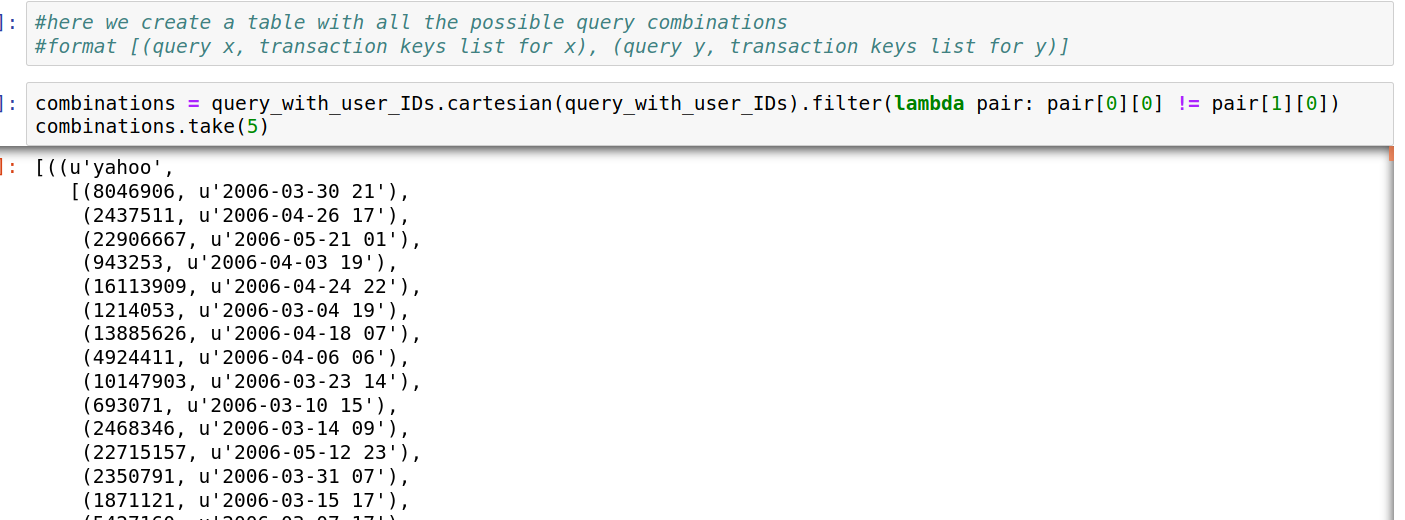
שיהיה לנו מבנה יותר מתאים לחשב את הקונפידנס הרצוי לכל קומבינציה

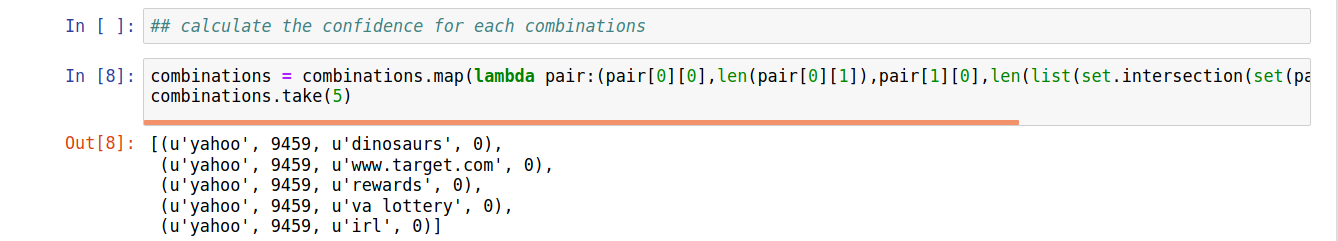
בכדי לחשב את הקונפידנס נשים לב כי :

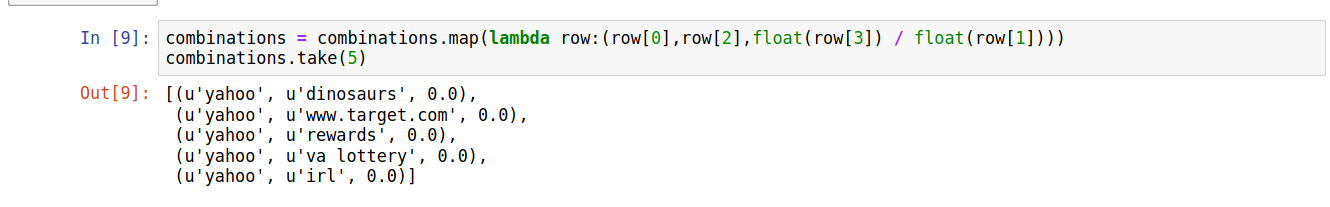


לכן חישבנו עבור כל קומבינציה של שתי שאילתות את מספר הטרנזקציות שמכילות את X, מספר הטרנזקציות שמכילות את y , מספר הטרנזקציות שמכילות את X ו Y

ולאחר מכן כמו המשוואה חישבנו בפשטות את conf(x=>y)



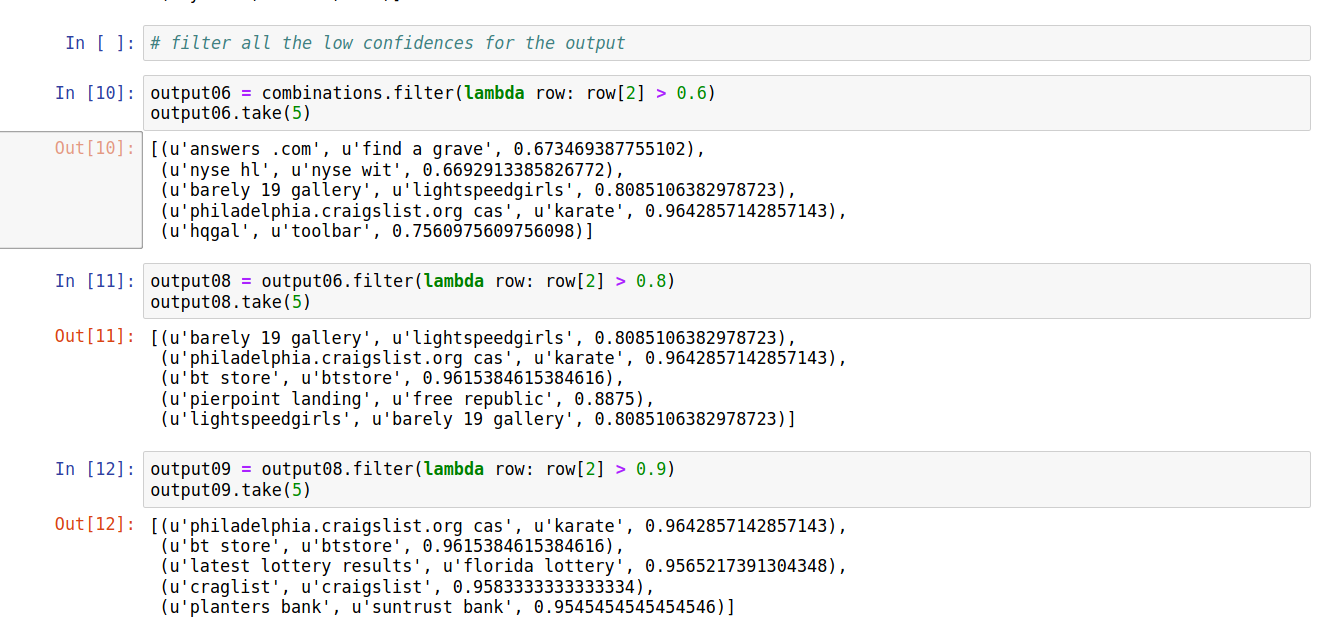




מהתמונה השנייה נראה קונפידנס נמוך ,כלומר אין קשר מהותי בין yahoo לtarget .

**שלב 7:**

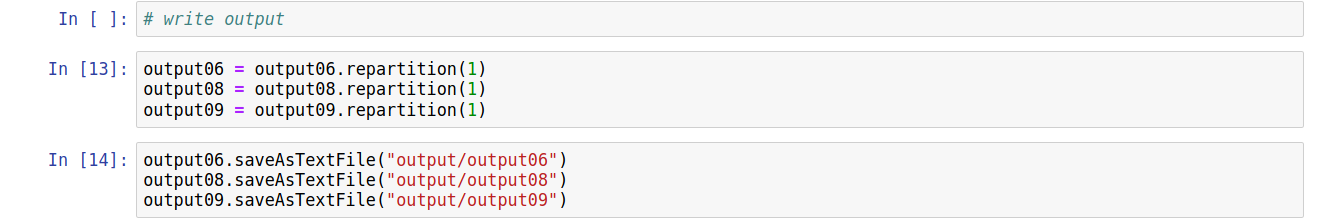
כעת נרצה להציג כפלט את הקונפידנס הנדרש בתרגיל כפלט.ראשית מהטבלה שנוצרה לנו בשלב הקודם נשאיר תחילה את אלו שהקונפינדס שלהם מעל 0.6 ולאחר מכן 0.8 ובסוף 0.9



**שלב 8 (שלב אחרון)**

נדפיס את תוצאות לתוך קבצי טקסט,

הקבצים מצורפים בקובץ ההגשה.



**גרירות מעניינות שמצאנו:**

**planters bank => suntrust bank 0.9545454545454546**

מצאנו כאן קשר חזק בין שני הבנקים יתכן ש 95% מהלקוחות של מנוע החיפוש שיש להם לשבון ב planters bank יש להם גם חשבון ב suntrust bank

**lightspeedgirls => barely 19 gallery 0.8085106382978723**

החיפושים הללו הם מעולם הפורנו לכן נחסוך את הפירוט..

**latest lottery results => florida lottery 0.9565217391304348**

מהגרירה הזאת אנחנו יכולים להבין כי 95% מהלקוחות של מנוע החיפוש ששלחו לוטו ורצו לבדוק את התוצאה של ההגרלה האחרונה שלחו את טופס הלוטו ב florida lottery

**yeshivaworld => theunorthodoxjew.blogspot.com 0.8064516129032258**

yeshivaworld זהו אתר חדשות חרדי theunorthodoxjew.blogspot.com זהו אתר קהילה חרדי שמכיל הרבה מידע אודות רבנים וחומר לימודי