אנליזה של ביג דאטה – תרגיל 2:

זרימת האלגוריתם-

בשלב הראשון טענו את הdataSet ופירסרנו אותו ע"י הורדת הכותרות, פיצול לשורות לפי tabים והמרת עמודת הuserID לint על מנת שנוכל לעבוד עם המידע בצורה נוחה ומסודרת.

פילטרנו את המידע ע"י הורדת כפילויות של משתמשים שחיפשו את אותו ערך יותר מפעם אחת.

```
In [5]: filtered_dataset.take(10)

Out[5]: [(142, 'dfdf'),
         (142, 'vaniqa.comh'),
         (142, '207 ad2d 530'),
         (142, 'attornyleslie.com'),
         (217, 'mizuno.com'),
         (217, "p; .; p;' p; ' ;' ;';"),
         (217, 'yahoo.com'),
         (217, '-'),
         (1268, 'sstack.com'),
         (1268, 'www.raindanceexpress.com')]
```

את המידע המפולטר המרנו לdataFrame, ובאמצעות אגריגציה יצרנו מנבה נתונים בו כל שורה מכילה שאילתה ורשימה של כל המשתמשים שבצעו אותה.

```
In [10]: q_more_than_2_user.take(10)

Out[10]: [Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 21806
840, 3107412, 5819277, 7402970, 7476529, 20693622]),
          Row(Query='.hotmail.comhttp', UserID=[3732132, 11990167, 18253475, 21398243]),
          Row(Query='.wamu.com', UserID=[16455679, 3313123, 24535160, 18447964]),
          Row(Query='2 flash games', UserID=[4031594, 3781322]),
          Row(Query='aau.com', UserID=[15336031, 6482128, 2286838]),
          Row(Query='acris', UserID=[1730018, 2272416, 4821757, 9456273, 6820867]),
          Row(Query='action village', UserID=[3144279, 21620700]),
          Row(Query='affordable health insurance', UserID=[307464, 1260412]),
          Row(Query='ako', UserID=[1039396, 3172266, 9640439, 1408653, 1753504, 2706422, 13855112, 18760097, 2496878, 158842
14, 18167739, 386728, 808930, 1334291, 2191663, 2276626, 2653464, 5933562, 11034304, 20471493, 1413474, 2595980, 29
26375, 6155544, 6418168, 9324718, 9596629, 22984027, 955503, 2327099, 3095767, 3294416, 3949075, 5610864, 11581380,
15422574, 17344636, 1010520, 2000488, 18910912, 20248817]),
          Row(Query='alfie soundtrack', UserID=[5011024, 6109712])]
```

בשלב זה הורדנו את כל השאילתות שבוצעו על ידי משתמש יחיד על מנת להוריד מידע שלא יועיל לנו למציאת קשרים בין חיפושים.

על טבלה זו הפעלנו את הפקודה cartesian ויצרנו מכפלה קרטזית על מנת ליצור זוגות של שאילתות עבורן ננסה למצוא קשר. במהלך ביצוע המכפלה ביצענו פילטר כך שלכל זוג תהיה שורה בודדת ובשורה זו נבדוק את הקשר בשני הכיוונים.

```
In [13]: q_more_than_2_user_ct.take(5)

Out[13]: [(Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 2180
6840, 3107412, 5819277, 7402970, 7476529, 20693622]),
           Row(Query='.hotmail.comhttp', UserID=[3732132, 11990167, 18253475, 21398243])),
          (Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 2180
6840, 3107412, 5819277, 7402970, 7476529, 20693622]),
           Row(Query='.wamu.com', UserID=[16455679, 3313123, 24535160, 18447964])),
          (Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 2180
6840, 3107412, 5819277, 7402970, 7476529, 20693622]),
           Row(Query='2 flash games', UserID=[4031594, 3781322])),
          (Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 2180
6840, 3107412, 5819277, 7402970, 7476529, 20693622]),
           Row(Query='aau.com', UserID=[15336031, 6482128, 2286838])),
          (Row(Query='...', UserID=[3554879, 4005384, 6928849, 19135358, 1812007, 9275494, 5226229, 13384381, 16941836, 2180
6840, 3107412, 5819277, 7402970, 7476529, 20693622]),
           Row(Query='acris', UserID=[1730018, 2272416, 4821757, 9456273, 6820867]))]
```

בנקודה זו יש בידנו טבלה בא כל שורה מייצגת זוג שאילתות ורשימת משתמשים עבור כל שאילתה.

האלגוריתם עובר שורה שורה בטבלה ומכניס אליה לשתי עמודות חדשות את הconfidence של X=>Y ושל Y=>X.

את הconfidence אנו מחשבים בפונקציה שמקבלת את המשתמשים שחיפשו את X ואת המשתמשים שחיפשו את Y. ומחלקת את אורך החיתוך שלהם באורך הרשימה של X.

```
In [16]: q_more_than_2_user_union.take(1)

Out[16]: [('...',
           [3554879,
            4005384,
            6928849,
            19135358,
            1812007,
            9275494,
            5226229,
            13384381,
            16941836,
            21806840,
            3107412,
            5819277,
            7402970,
            7476529,
            20693622],
           '.hotmail.comhttp',
           [3732132, 11990167, 18253475, 21398243],
           0.0,
           0.0)]
```

כעת עברנו על שורות הטבלה המלאה והסרנו ממנה כל שורה כך שהconfidence של הצמד אותו היא מייצגת שווה ל0.

```
In [19]: q_more_than_2_user_union_filter.take(1)

Out[19]: [('guess bags',
           [471426, 999001, 1799032],
           'tvguide',
           [1669898,
            3053569,
            6068638,
            7924610,
            12169244,
            354548,
            7583211,
            13220997,
            227785,
            5151588,
            2815747,
            1799032,
            2902572,
            16350598,
            16893187,
            4341955,
            471426,
            6001607],
           0.6666666666666666,
           0.1111111111111111)]
```

את הטבלה הדפסנו לקובץ טקסט.

חישוב לפי confidence מסוים (0.6, 0.8, 0.9)-

אנחנו מריצים את האלגוריתם 3 פעמים כך שבכל פעם נשנה בשלב הסינון את confidence ומדפיסים לקובץ.

קשרים מעניינים-

- (doing too much) => (paula deanda) 0.6666666666666666
- (paula deanda) => (doing too much) 0.2857142857142857

זמרת אמריקאית והביטוי 'עושה יותר מידי'

- (rob zombie) => (dani filth) 0.16666666666666666

שני זמרים מהרכבי מטאל

- (insane clown posse) => (dani filth) 0.2222222222222222

זמר מטאל וביטוי 'תנוחת ליצן משוגע'

- (law and order svu) => (mariska hargitay) 0.5

שם של סדרת טלויזיה ושחקנית שמשחקת בסדרה

- (kingdom hearts2) => (naruto) 0.5

סדר וסרט אנימה