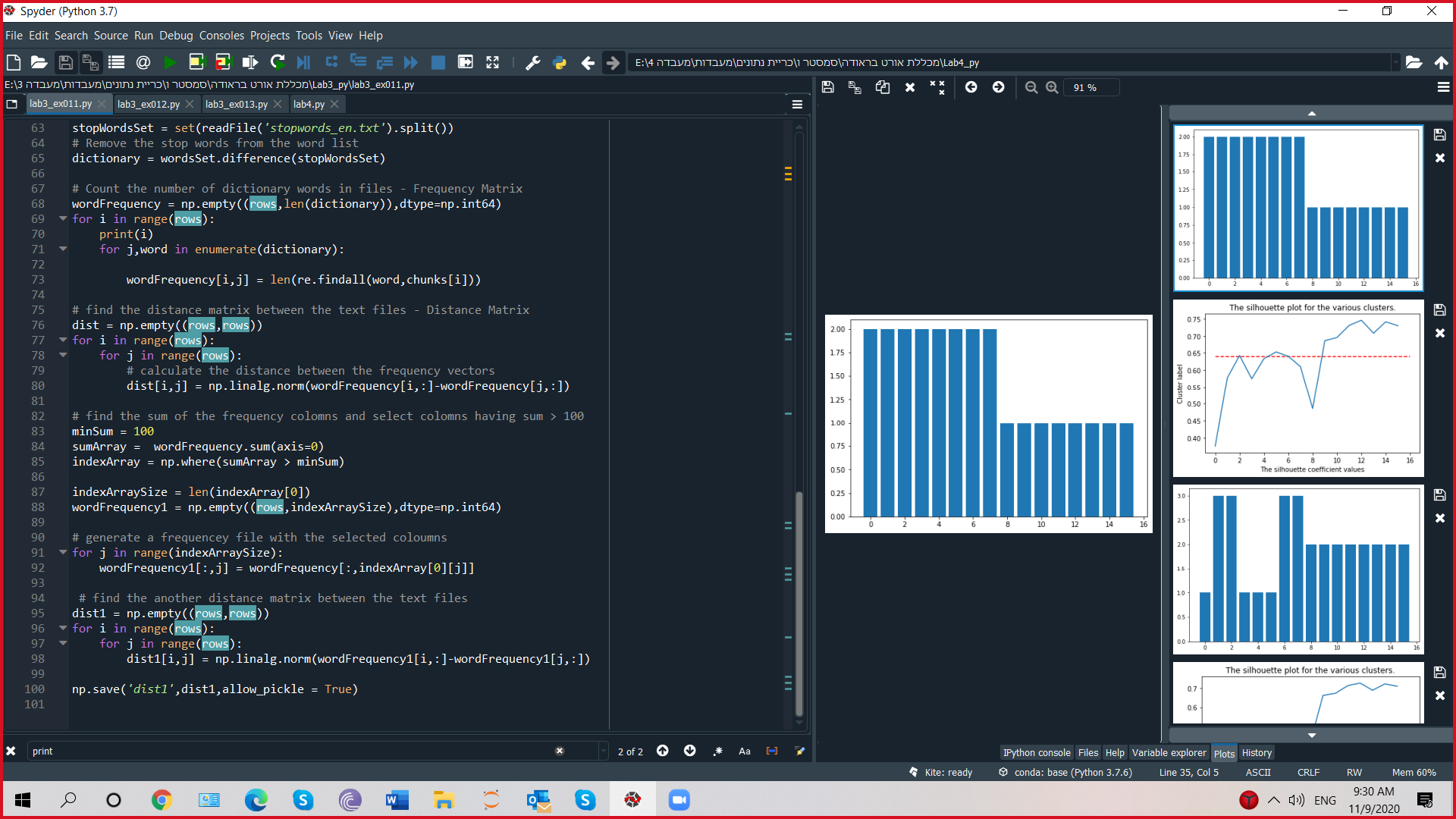
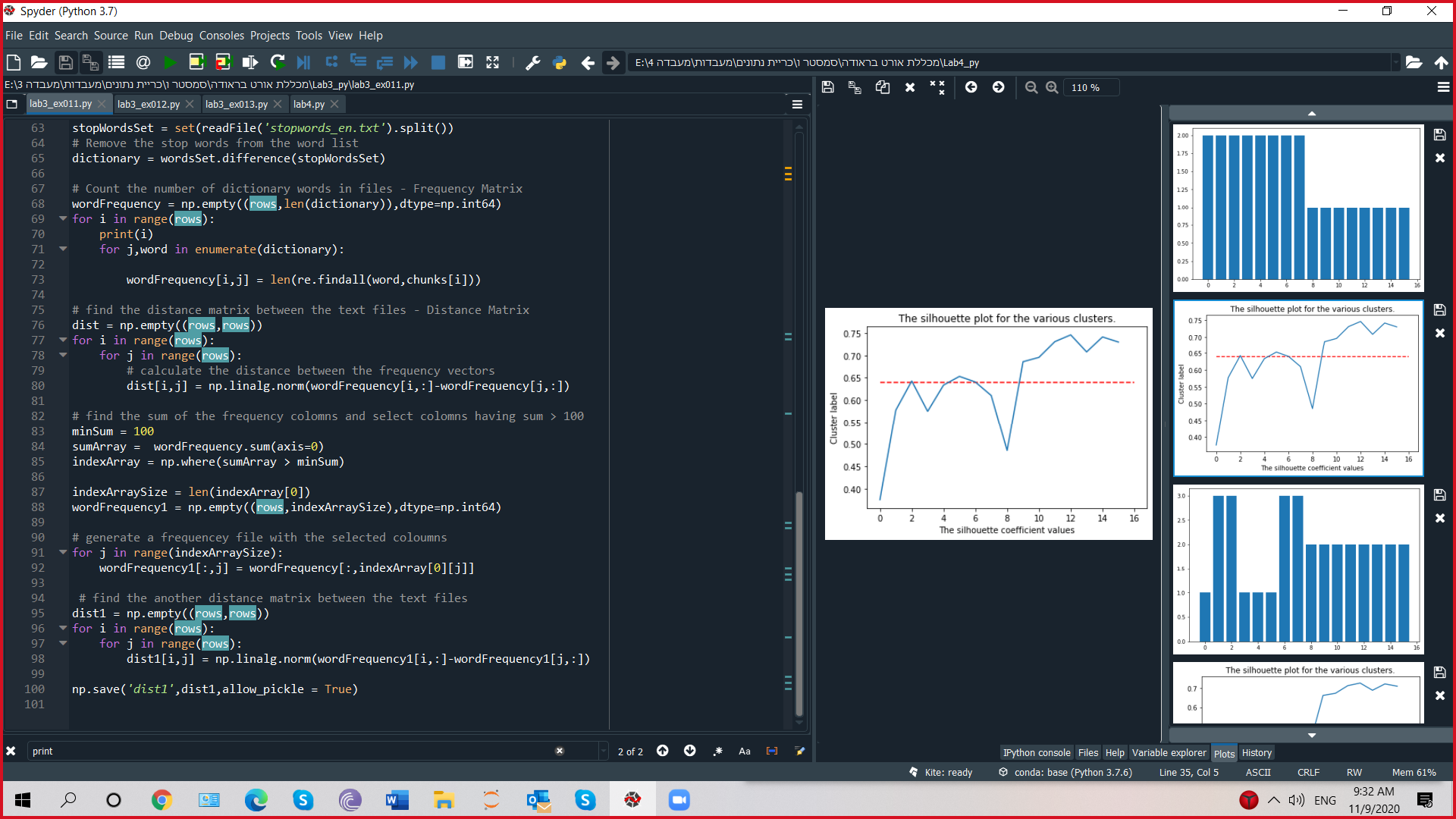
**מעבדה 3 – כריית נתונים**

שם מגיש: אילון בן סימון ת"ז: 312162951

שם מגיש: סער ויקטור ת"ז: 312392822

**תחילה בדקנו עבור wind =20000:**

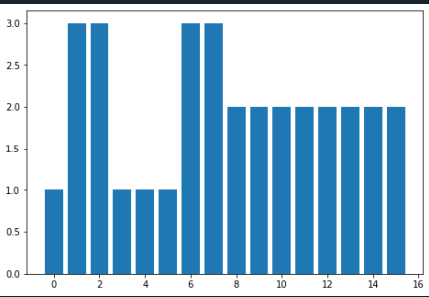
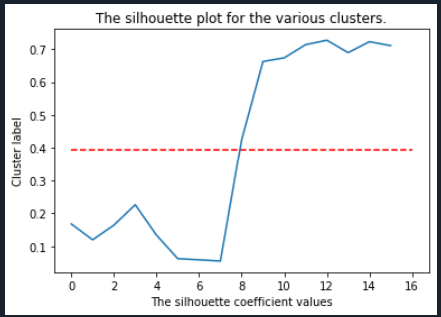
עבור חלוקה ל- cluster=2:



כאשר מחלקים ל-2 clusters ערך הסילואט הממוצע הינו:

. 0.6394586475962402

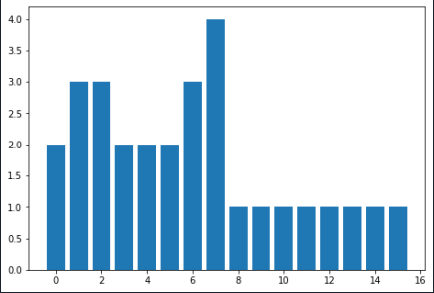
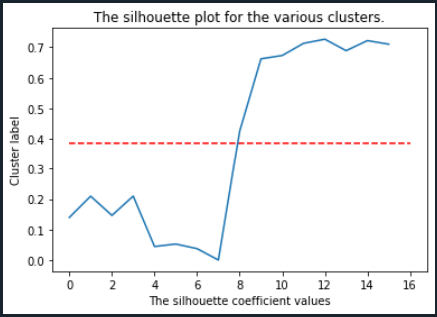
עבור חלוקה ל- cluster=3:



כאשר מחלקים ל-3 clusters ערך הסילואט הממוצע הינו:

. 0.3944732644738571

עבור חלוקה ל- cluster=4:



כאשר מחלקים ל-4 clusters ערך הסילואט הממוצע הינו:

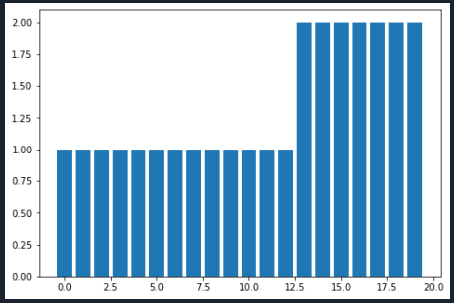
0.3849549157203973.

מסקנות ניסוי wind = 50,000:

ניתן לראות שכאשר החלוקה היא לשני clusters אנחנו מקבלים את ערך הסילואט הממוצע הגבוה ביותר (0.639) בפער גדול מהאחרים (0.38 ו -0.39) כלומר ניתן להסיק מכך שקיימים 2 ספרים.

לאחר שבצענו חלוקה לשני ספרים, הרצנו את האלגוריתם שוב אך הפעם עם wind קטן יותר (5000) ועל שני הצ'אנקים שנמצאו באינדקסים 7 ו-8 (היכן שניתן לראות שספר 1 נגמר וספר 2 התחיל) וקיבלנו את התוצאות הבאות:





מצאנו שהספרים מתמזגים בין הצ'אנקים שנמצאים באינדקסים 12 ו-13 כלומר נקודת המיזוג בחלק היחסי שדגמנו (100,000 בתים) היא בין 60,000 בתים ל- 65,000 בתים. וביחס לכל הקובץ המקורי נקודת המיזוג היא בין 410,000 בתים ל-415,000 בתים.

להלן הפונקציה למציאת גבולות של הספרים בצורה אוטומטית:

def detectBoundaries(label):

for i in range(0, len(label)-1):

if label[i] != label[i+1]:

end\_book\_1 = i

start\_book\_2 = i + 1

return end\_book\_1, start\_book\_2

return -1, -1