**Assignment 3 - Report**

השתמשנו ב Jupiter Notebook על מנת לעזור לנו לתכנן סידרה של בדיקות על המידע.

ניקח צילומי מסך משם לצורך הדוח שלנו ולבסוף, לצורך הסדר, נשאיר בקובץ ה Lab3.py רק את המודל המוצלח שלנו.

(השארנו כפונקציות את שאר המודלים במידה ויש צורך לבדוק אותם [המודלים האלה נלקחו מה Jupyter notebook שלנו])

חשוב לציין כי עשינו עוד הרבה ניסויים עם כל מיני מודלים אך למען הסדר בחרנו לקחת את אלה עם התוצאות הטובות ביותר ולהציג בדוח.

השם שבחרנו לתחרות ב Kaggle הוא Saar Guttman

ראשית טענו את ה data.

התחלנו בלעבור על המידע ומיד הבנו שצריך בזמן ה preprocessing להשמיט את כל התיוגים (@<name>) בגלל שאי אפשר לקשר התיוגים לחיוביות של המשפט.

בנוסף ראינו כל מיני דוגמאות של אימוג'ים הבנו שכדאי לעבור עליהם ולהחליף את ה אימוג'י בשם

לדוגמה:

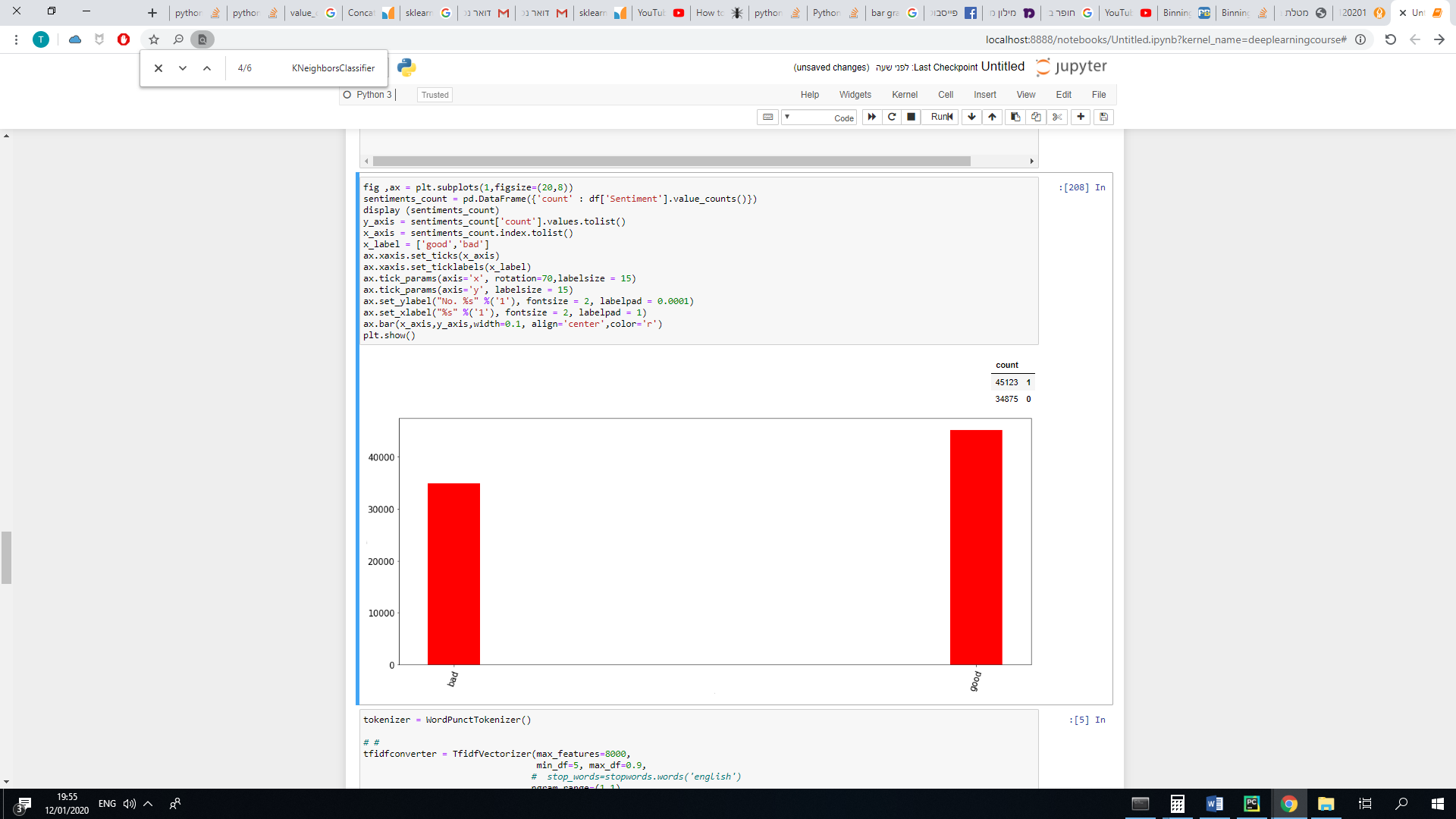
על מנת ליצור קשר בין המשפטים, בגלל שבדרך כלל כאשר המשפט הוא שלילי אנשים אולי יוספו אימוג'י עצוב ( )

לאחר מכן גם נתקלנו גם ב Sentimet-Text שהם לינקים לאתרים אחרים והחלטנו לסנן גם אותם.

1. ראשית בדקנו את ההסתברות להיות בכל אחד מה class-ים

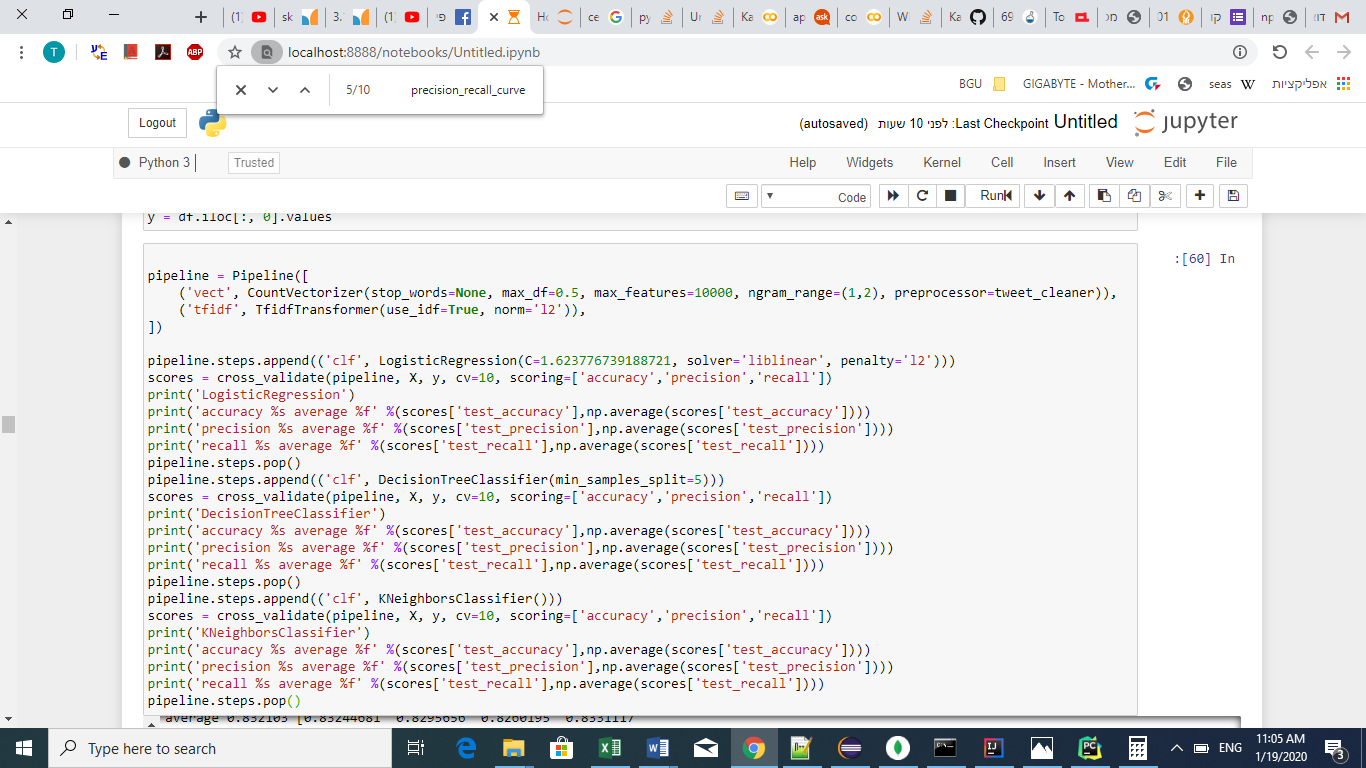
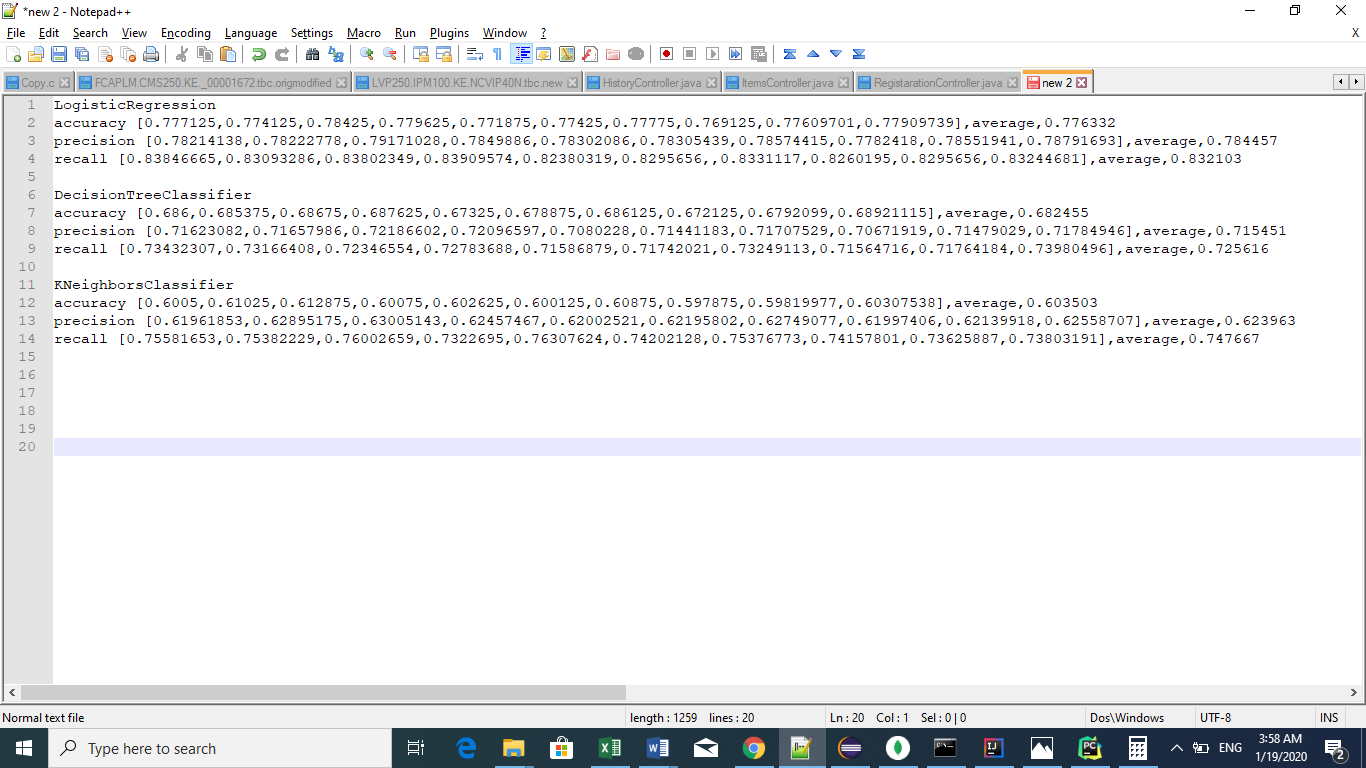
כפי שניתן לראות בתמונה, הסיכוי להיות שייך ל:

* 0(משפט שלילי) :
* 1 (משפט חיובי) :



רצינו כי התוצאה הסופית שלנו תהיה גבוהה מתוצאת הניחוש

1. התחלנו בלהכין tfidf פשוט ואלו התוצאות שקיבלנו

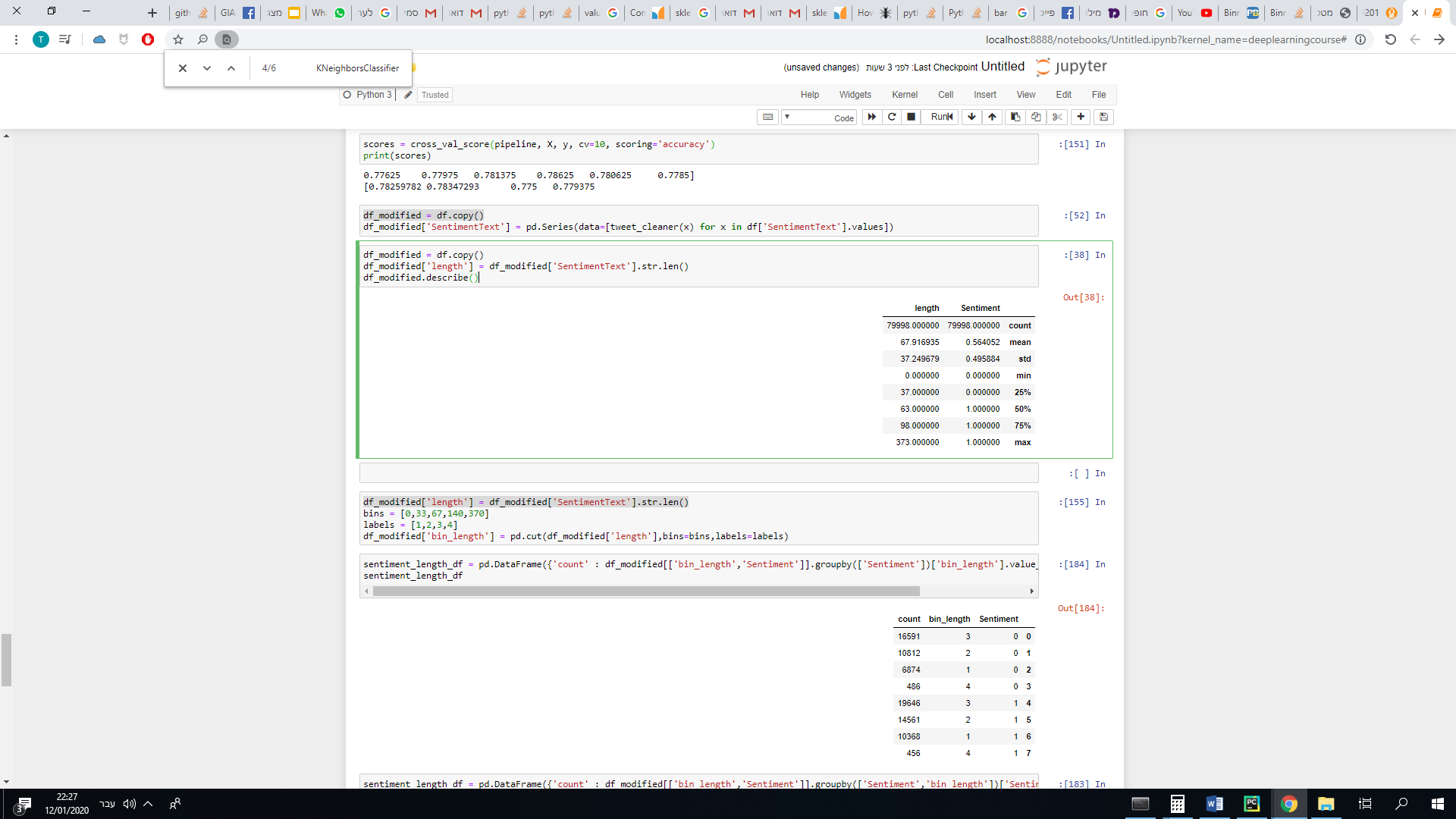


ממוצע Accuracy של

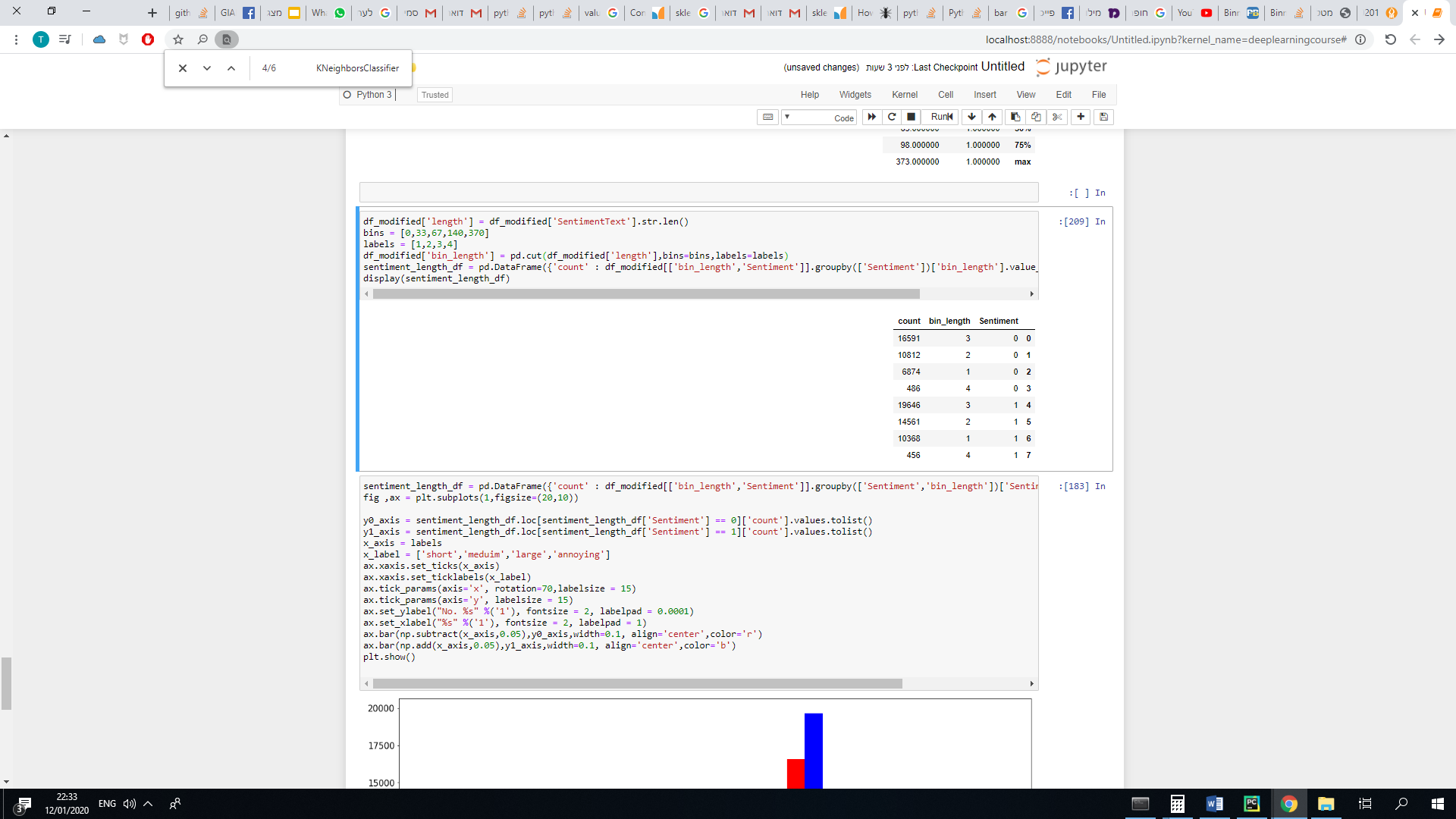
הרבה יותר טוב מניחוש!

1. בהמשך המשכנו להסתכל על ה data והעלנו השארה כי כמות התווים במשפט עשויה להיות קשורה לחיוביות שלו.

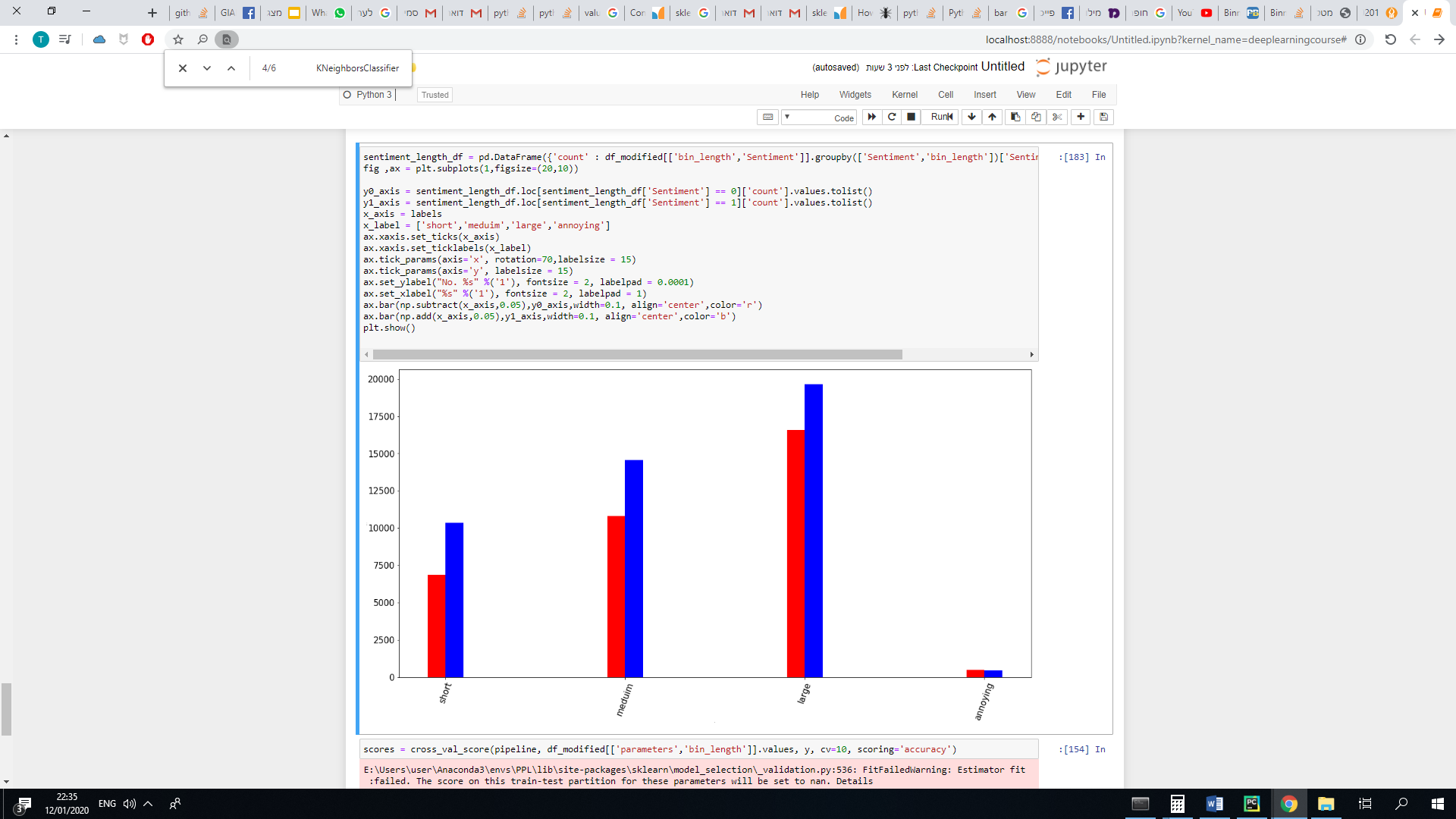
לצורך הבדיקה הזאת עמודה לדשה ב DataFrame הנקראת length ובדקנו מה הוא הרבעון, החציון וכדו..



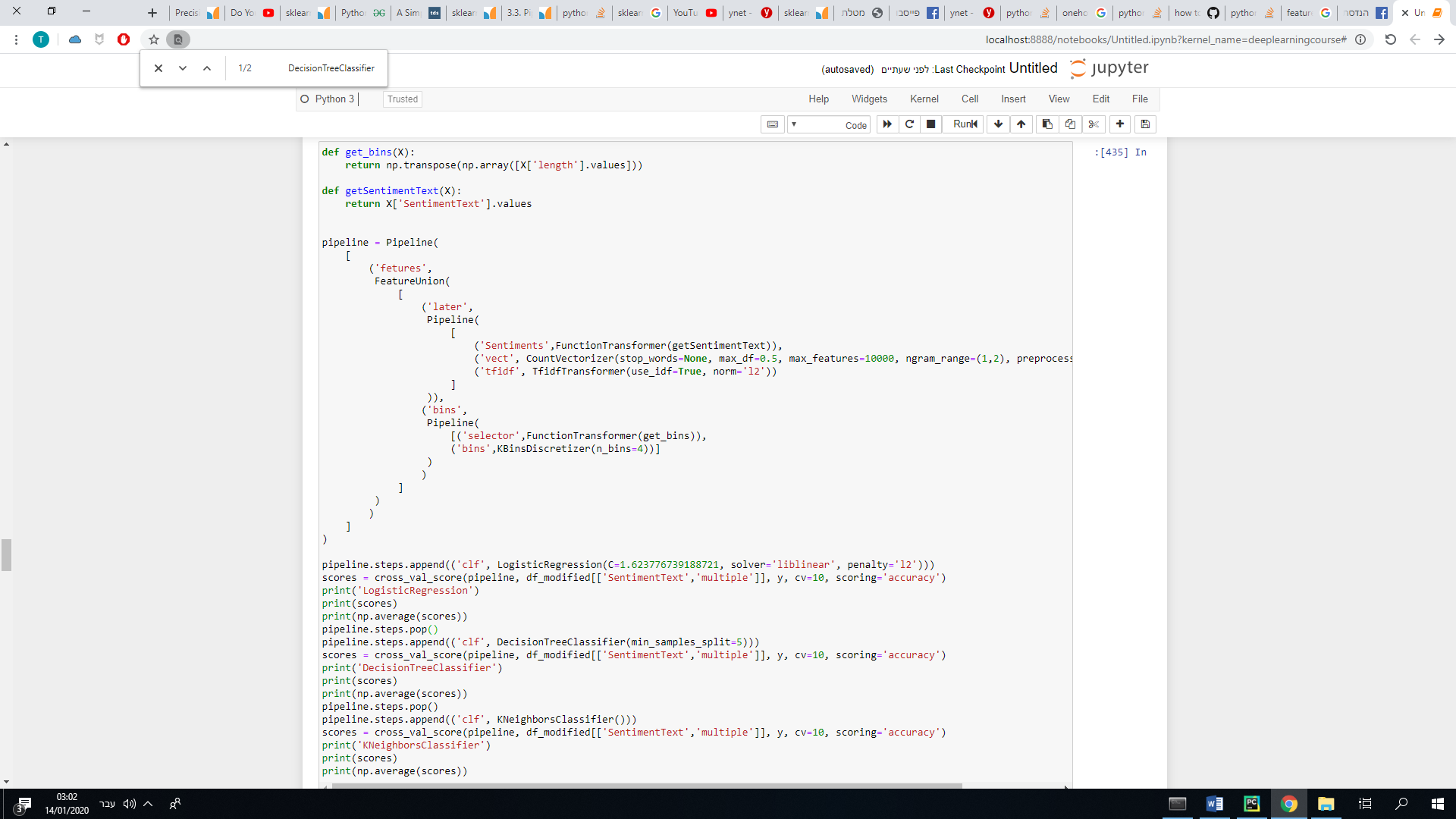
על בסיס הנתונים האלה ועל בסיס המידע שלנו על גודל ציוץ מקסימאלי החלטנו לפצל ל 4 bins בצורה הבאה

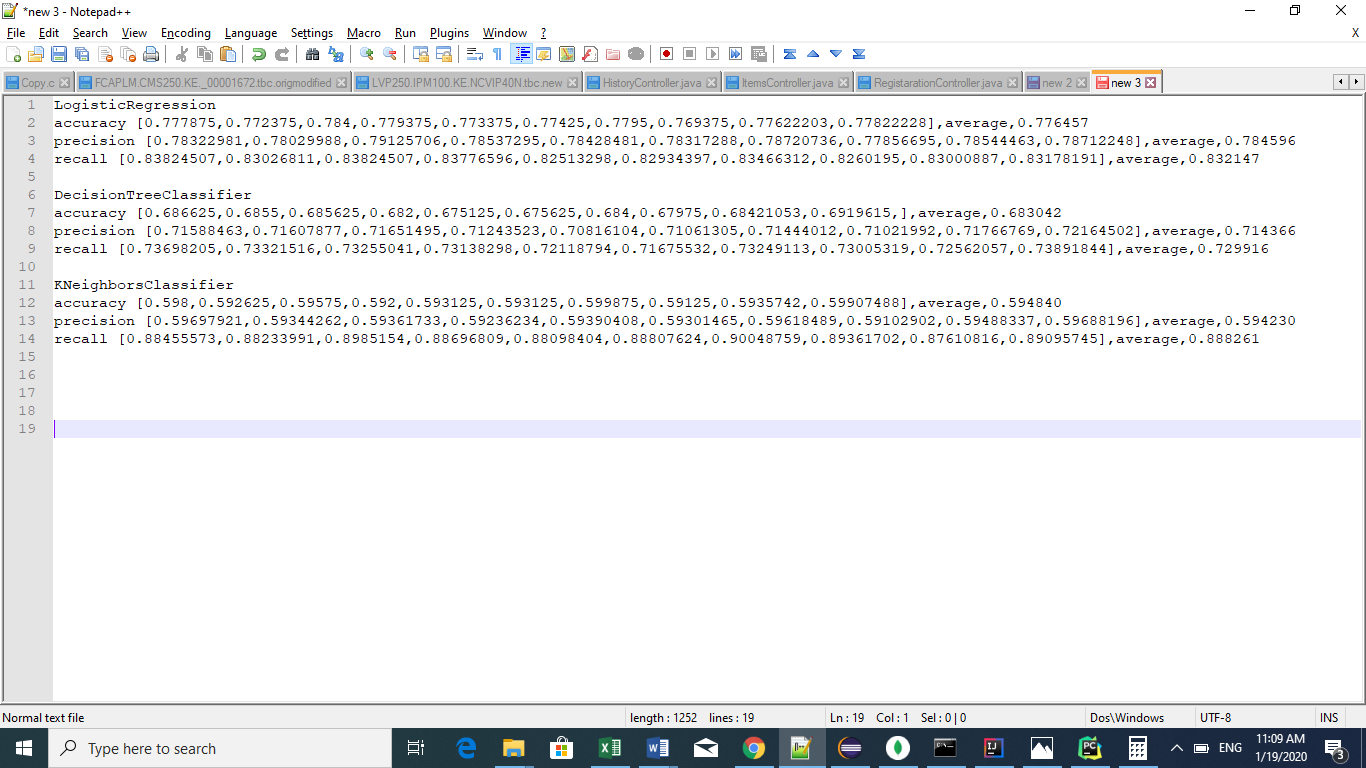


העלנו את הנתונים האלה על גרף כך שמשפט חיובי יסומן בכחול ומשפט שלילי יסומן באדום



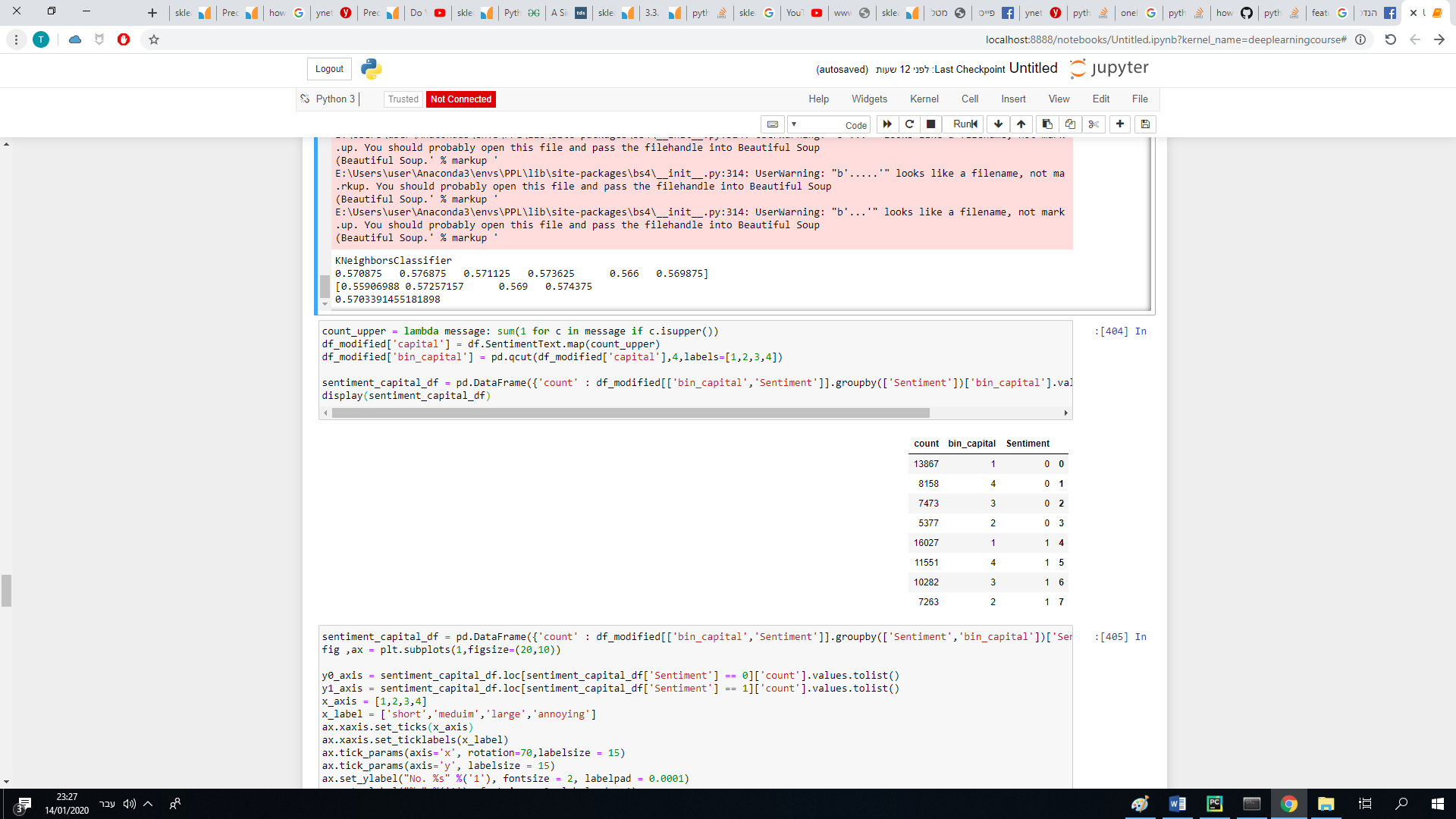
כפי שניתן לראות, באופן כללי רוב המשפטים בכל הקטגוריות היו חיוביים, מלבד שפטים ממש ערוכים ניסנו להוסיף את ה feature הזה אך לצערנו הוא לא שיפר את התוצאות



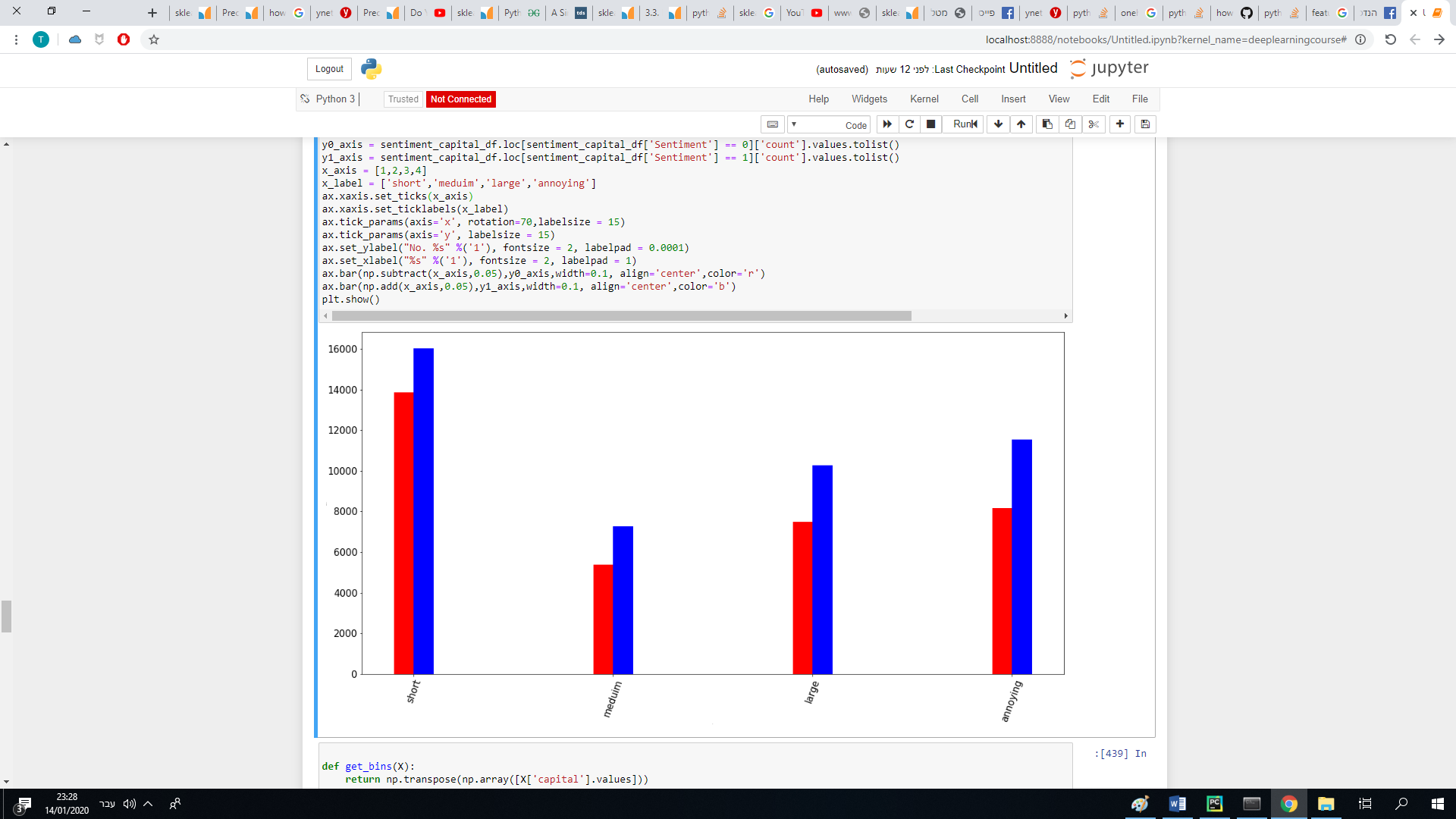


לא הצלחנו להגיע לשיפור בתוצאות מהמודל הראשון.

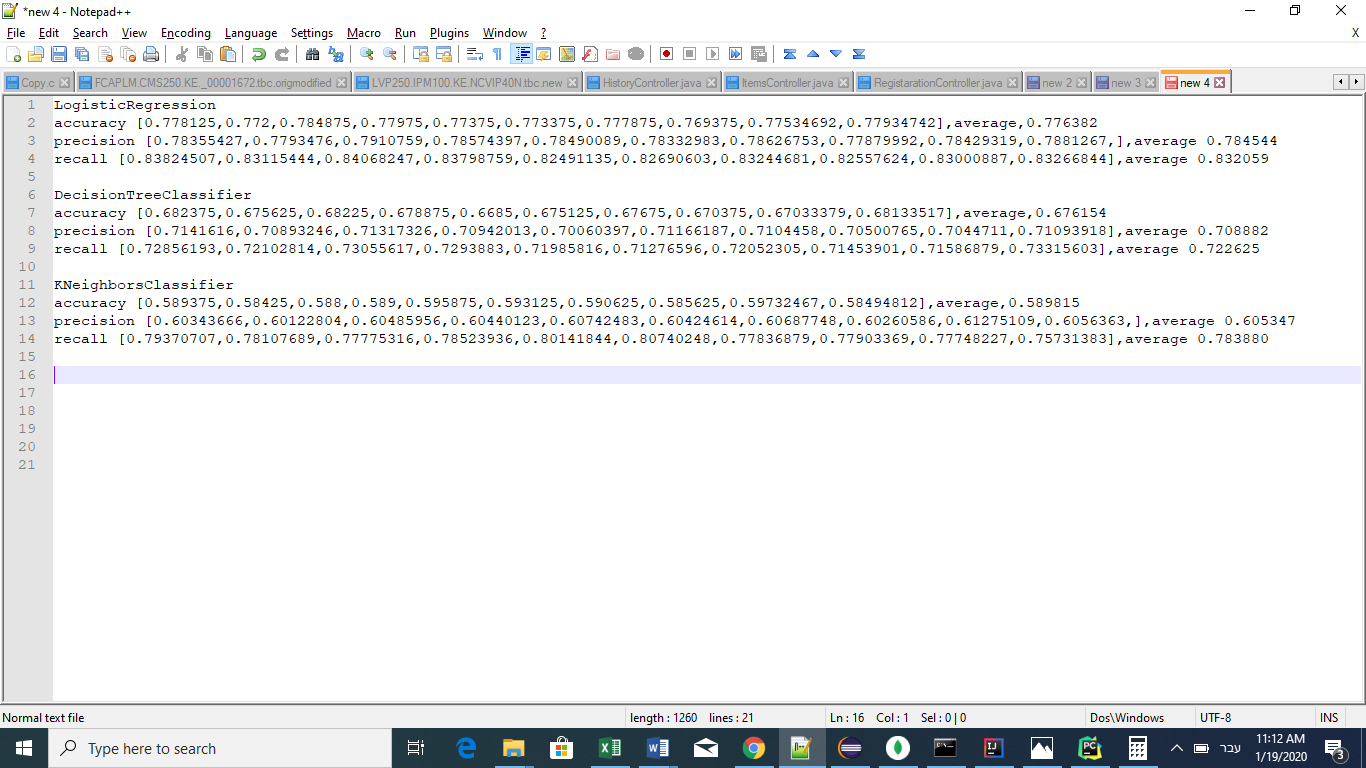
1. לאחר מכן ניסנו לבדוק אולי כמות האותיות הגדולות במשפט יוכל לעזור לנו להחליט על החיוביות שלו אז ערכנו עוד ניסוי



גם כאן פיצלנו ל 4 class-ים שונים



ואלו התוצאות



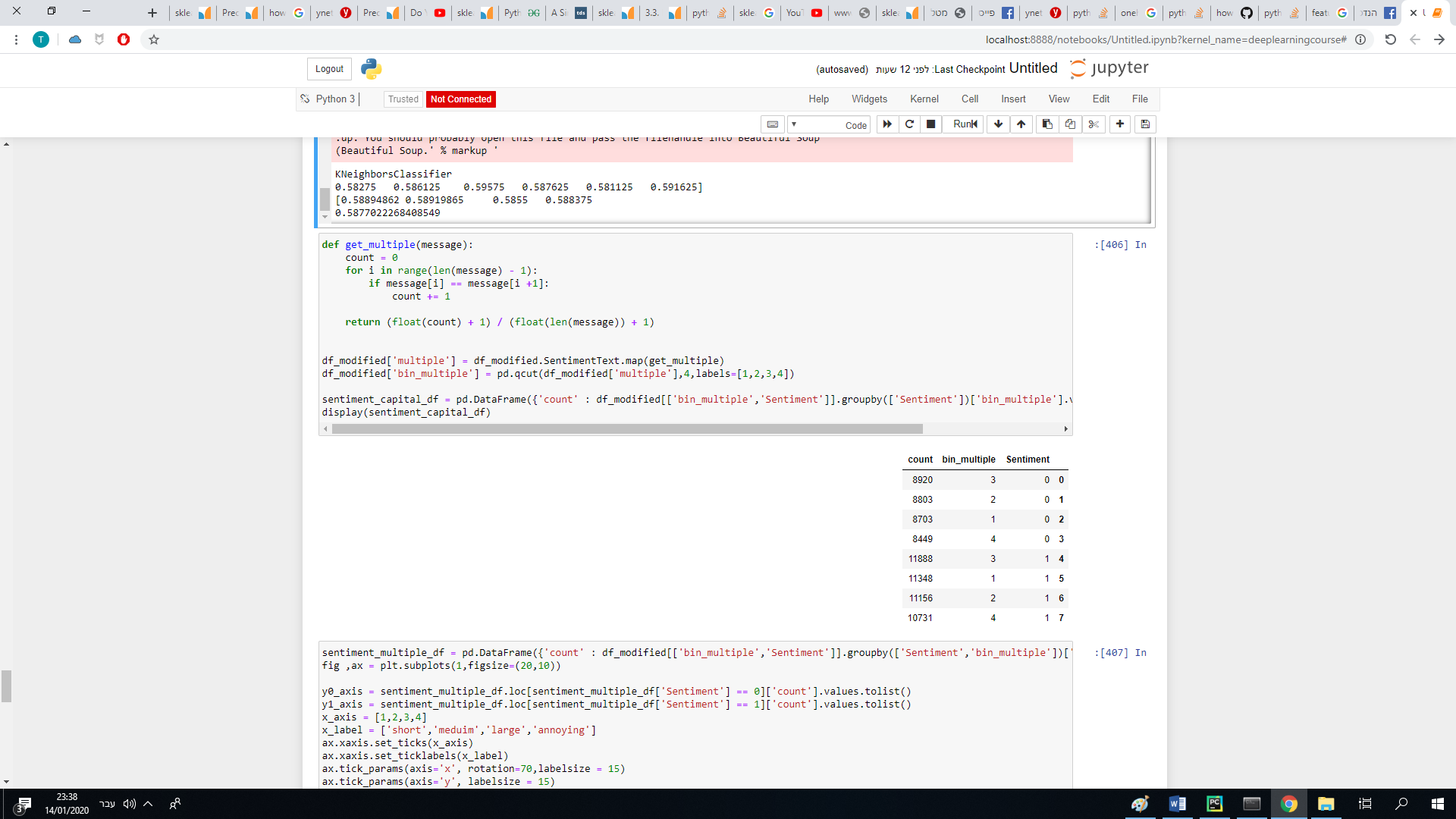
לא ראינו שיפור מהמודל הראשון והשני וניסינו בדרך אחרת.

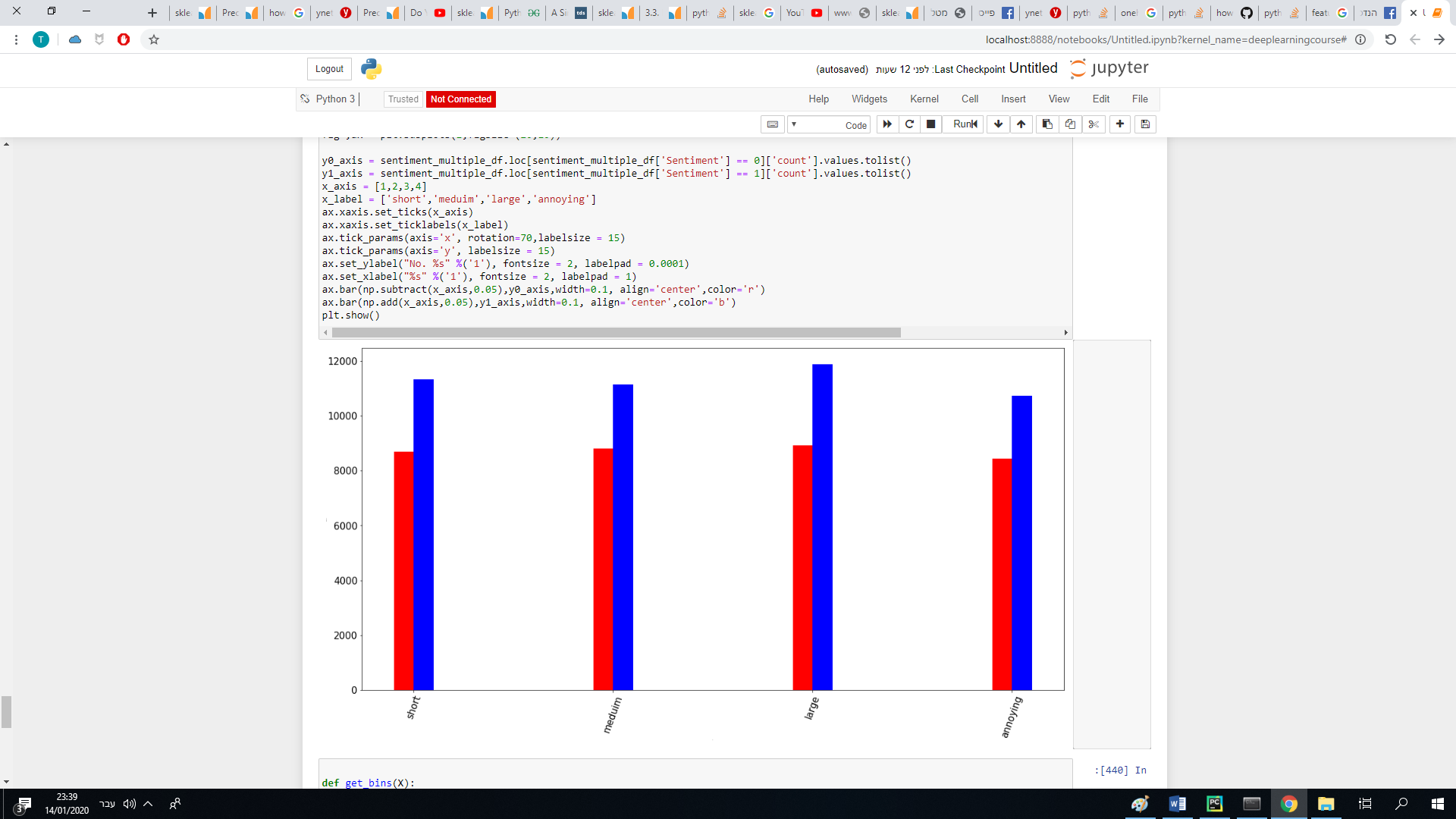
1. לאחר מכן בדקנו את היחס בין כמות האותיות במשפט לכמות הפעמים שאות חוזרת על עצמה אחת אחרי השנייה

לדוגמה:

* Goooallllllllllllllllllllllllllll
* Yeahhhhhhhhhhhh
* Shiiiiiiiiiiiiittttttttttttttt

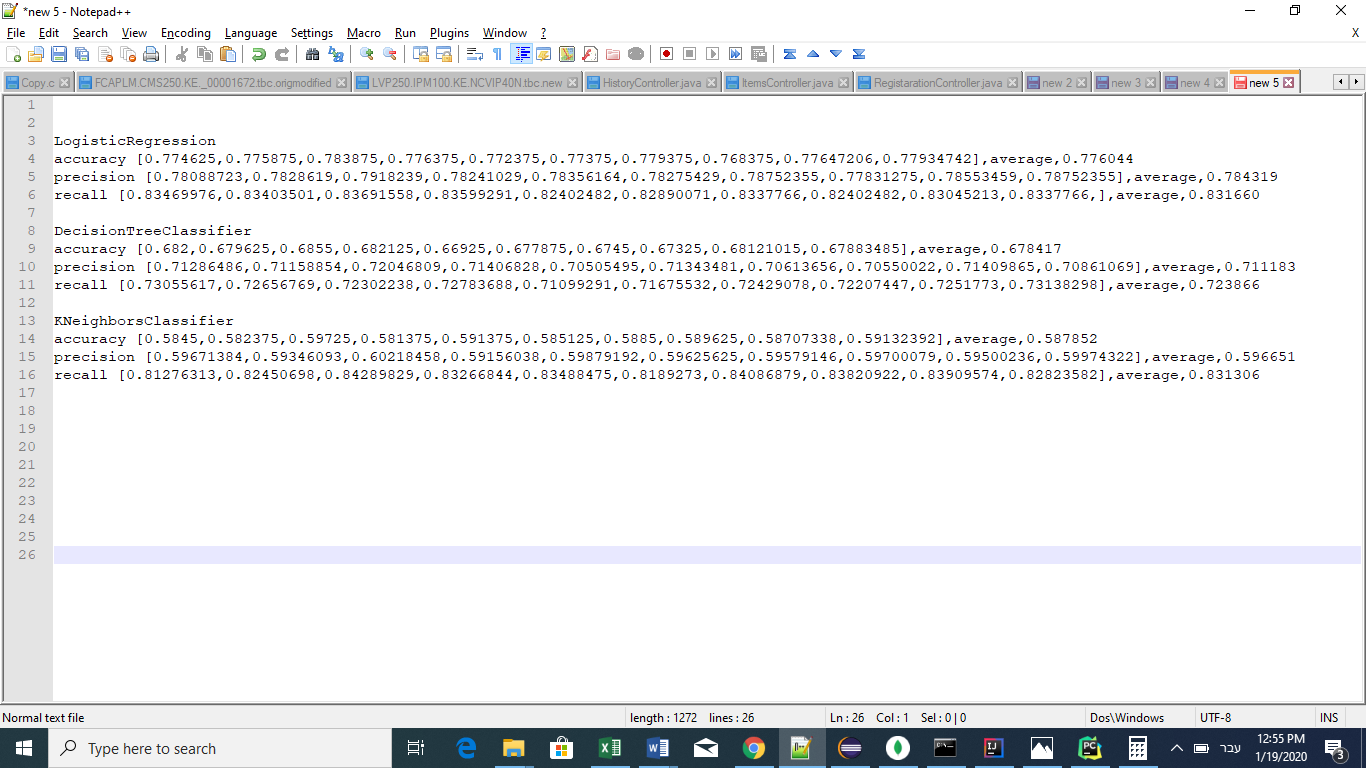
ורצינו לבדוק האם זה יכול לעזור בהחלטה על החיוביות של המשפט, גם כאן חילקנו לקבוצות





שוב ראינו כי הרוב מכיל משפטים חיוביים.

בנינו מודל חדש עם הנתונים החדשים הרצנו ואלו התוצאות:



שוב לא היה שיפור.

1. החלטנו להשתמש במודל הראשון על מנת והעלות אותו ל Kaggle, הסיבה שבחרנו דווקא בו היא כי הוא פשוט ומספק ציונים טובים.

בניגוד לשאר המודלים שלנו בגלל הפשטות שלו הוא רץ בזמן סביר עבור 10-cross validation והציון שלו זהה ואפילו גבוהה מרוב המודלים שחשבנו שישפרו את התוצאה.