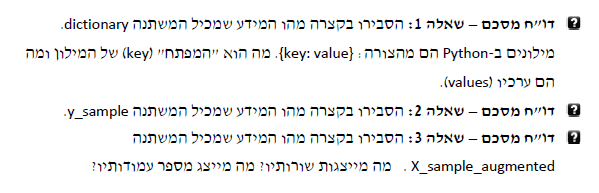
**מעבדה מערכות לומדות**

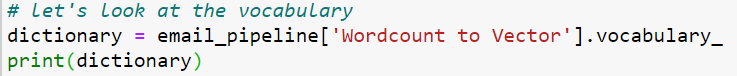
**חלק א**

מגישים:

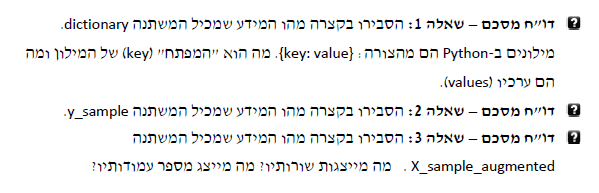
נתנאל רוטשילד 204937841

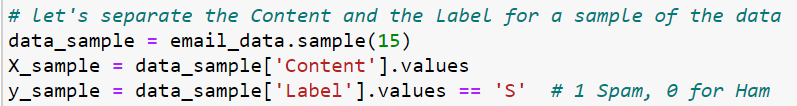
גיא פרג 201096534





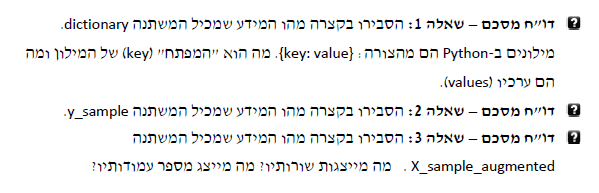
המשתנה dictionary מממש מילון אשר ערכי המפתחות הן המילים עצמם וערך כל מילה במילון היינו מספר אשר מייצג את מיקום אותה המילה במטריצה.



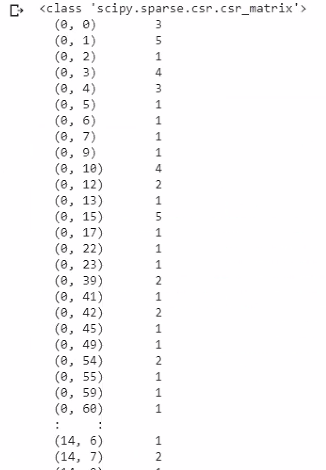


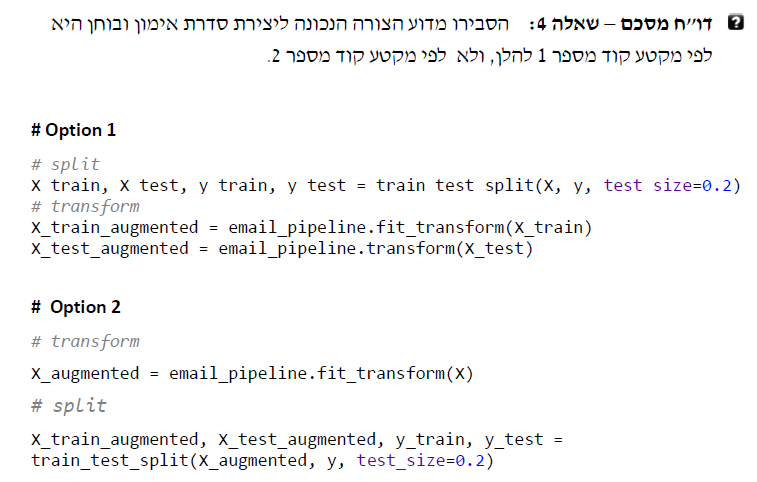
y\_sample מכיל את התיוג הנכון עבור כל אימייל.

הערך הבינארי לכל מייל מייצג בהתאמה # 1 Spam, 0 for Ham



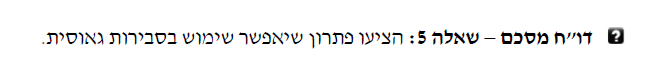
המטריצה מייצגת בתור השורות את המיילים את המילים כעמודות, כך שעבור כל צמד של (מייל, מילה) נקבל את ערכו המייצג את מספר המופעים של המילה במייל זה.



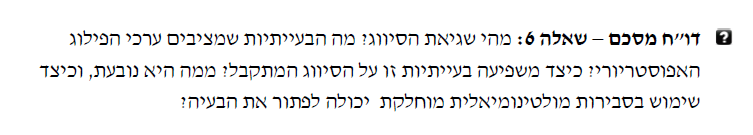


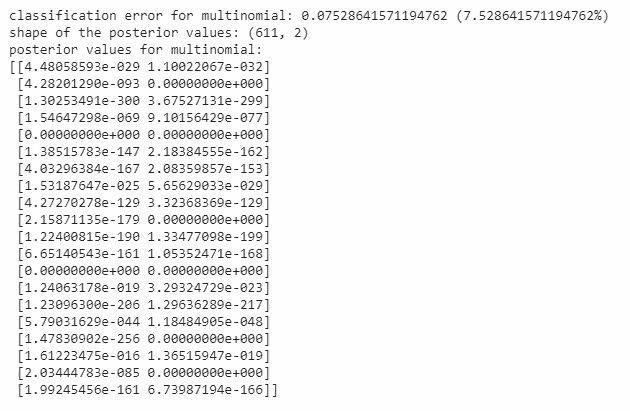
נשים לב להבדל בין קטעי הקוד: בקטע מספר 1 קודם מתבצע הפיצול לסדרות אימון ובוחן ורק לאחר מכן הטרנספורמציה (יצירת המילון) בניגוד למה שקורה בקוד 2 שם קודם בונים את המילון ורק לאחר מכן מבצעים את הפיצול. כלומר מתבצעת התעסקות עם סדרת הבוחן.

כמו שלמדנו בהכנה, הדרך הנכונה ליצירת סדרת אימון ובוחן היא זאת שלא תחשוף כלל את סדרת הבוחן למסווג בשום שלב מלבד שלב הבדיקה.



בשלב הראשוני נבצע בדיקה האם ישנן מילים המופיעות בכל המיילים בכמות זהה וגורמות ל-var=0, את אותן מילים נמחק בשלב הסינון הראשוני.



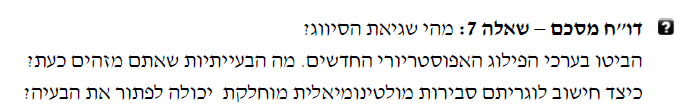


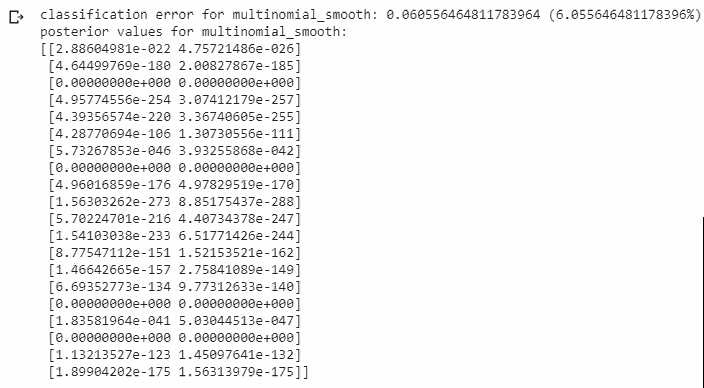
קיבלנו שגיאת סיווג של 7.52% .

ניתן לראות כי חלק נכבד מערכי הפילוג שמתקבלים הינם אפסים.

התאפסות זאת גורמת להורדת הסיכוי של דגימות חדשות לשיוך באחת המחלקות (אמת או ספאם), ולכן שיערוך מיקום הדגימה אנו נותן תוצאה ראלית.

שימוש בהסתברות מולטינומיאלית מוחלקת מאפשרת לנו לפתור בעיה זאת בכך שבמקרים בהם היינו מקבלים אפס עכשיו נוסף קבוע C (בדרך כלל ההסתברות הקטנה ביתר שהיא לא אפס) וכך לא יתקבל אפסים.

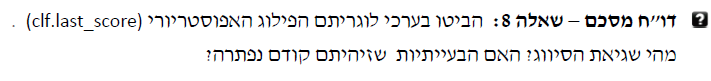




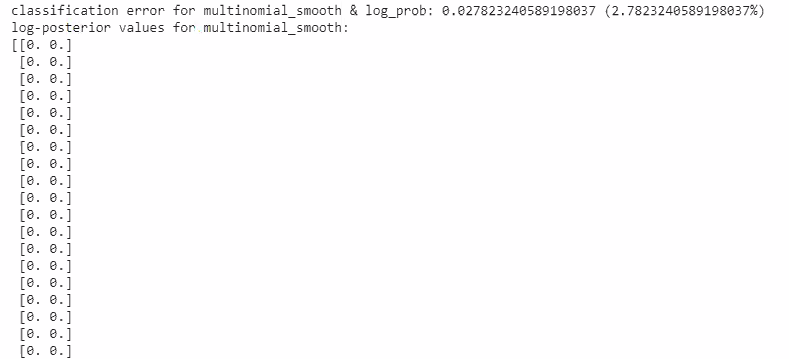
הפעם קיבלנו שגיאת סיווג של 6.05%

נשים לב כי אנחנו עדיין מקבלים אפסים בחלק מן הפילוגים, שכן גם הערך שהוספנו (כחלק מההחלקה) עדיין לא מספיק גדול כדי שהמחשב לא יחשיב את הפילוג בתור אפס.

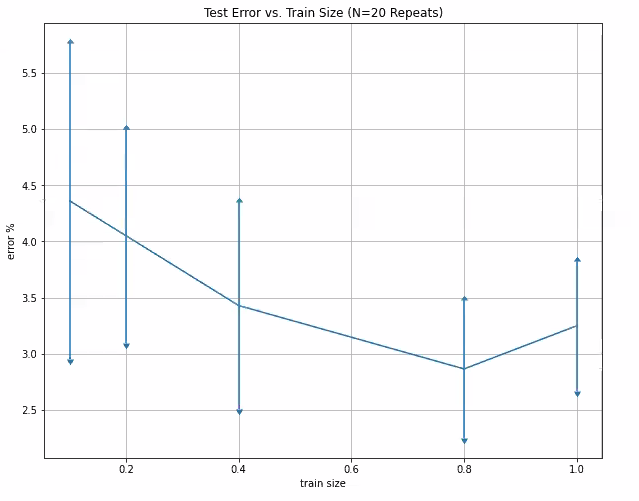
בפרט התעסקות עם הכפלה ומספרים קטנים מקשים על חישובי המחשב ולכן חישוב לוגריתם סיווג מולטינומיאלית מוחלקת יכול לתת לנו פתרון, גם מעבר מכפל לחיבור וגם הגדלת הערכים עליהם מתבצעים החישובים.



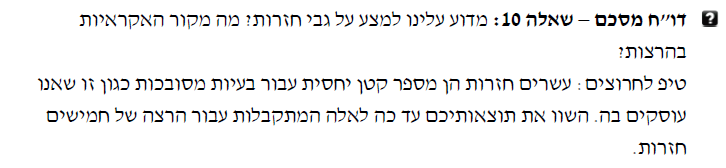
הפעם קיבלנו שגיאת סיווג של 2.78% , ואכן הבעיה נפתרה שכן קיבלנו ערכים גדולים יותר ולא קיבלנו אפסים.



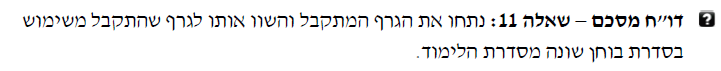




ניתן לראות כי קיבלנו את השגיאה הקטנה ביותר עבור 0.8 train size



קיבלנו שגיאות שונות עבור כל חזרה (בעקבות בחירות רנדומליות שונות של סדרות האימון והבוחן) ולכן יש לבצע ממוצע על כל השגיאות שנקבל.



ראשית נשים לב כי פיזור השגיאה כאשר קבוצת הבוחן מוכלת בקבוצת האימון קטן ביחס למצב בו קבוצת הבוחן וכקבוצת האימון זרות.

כאשר קבוצת הבוחן תפקדה גם כחלק מקבוצת האימון קיבלנו כי ישנה תלות בגודל הקבוצה.

עבור קבוצת אימון קטנה נקבל אחוז שגיאה ממוצעת קטן יותר, ועבור קבוצת אימון גדולה קיבלנו אחוז שגיאה קרוב לזה שקיבלנו עבור המצב בו קבוצת הבוחן וכקבוצת האימון היו זרות.

