**היבטים מעשיים בכריית תוכן אינטרנטי לאפליקציות עסקיות – פרויקט סיום**

**נושא הפרויקט:** Stock Market Prediction

**שאלה 1 – Data Exploration**

* השלב הראשון שהיינו צריכים לבצע על-מנת data exploration הוא לנקות את ה-data, בכדי שיהיה מנורמל:

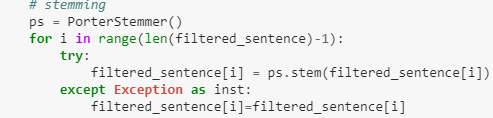
1. ביצוע Lowercase עבור כל מילה בטקסט:



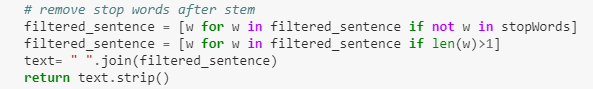
1. ביצוע Tokenizing עבור הטקסט (הפרדה למילים בצורת טוקן):



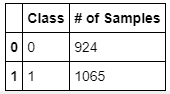
1. ביצוע Stemming עבור כל טוקן, תוך שימוש באלגוריתם porter (לא השתמשנו באלגוריתם Lemmatization מכיוון שלא רצינו להביא כמה מילים עם אותה משמעות, לאותה צורה. כלומר, רצינו לשמר את המילה המקורית שנכתבה בטקסט ורק לנרמל אותה לצורה הבסיסית שלה, ע"י "חיתוך" הסוף שלה.



1. הורדת Stop Words מן הטקסט, בכדי להתעלם ממילים נפוצות מידי בשפה האנגלית, אשר עלולות "להשתלט" על הטקסט וכתוצאה מכך להשפיע על התפלגויות המילים האחרות בידיעות (אותן אנו רוצים לחקור).



* לאחר ניקוי הנתונים, בדקנו אם הסיווג הבינארי הנתון הינו מאוזן/לא מאוזן (כלומר, אם רוב מוחץ של הרשומות שייכות לסיווג מסוים – 0 או 1). לכן, בדקנו כמה רשומות יש עבור כל סיווג:



ניתן לראות כי התפלגות הרשומות עבור כל Class הינה תקינה ומאוזנת (כלומר, אין מספר רב מידי של דגימות שמדד DJIA Adj עלה/ירד עבורן).

* בדיקת תדירויות 12 ה-terms הכי נפוצים בכל אחת משתי המחלקות, **בשני האופנים:**

1. עבור כל class, בדקנו עבור כל term **בכמה רשומות** הוא מופיע (כלומר, אם הופיע באותה רשומה כמה פעמים אז ספרנו זאת כפעם אחת בלבד לאותה רשומה).
2. עבור כל class, בדקנו **כמה פעמים** מופיע ה- .term

**שאלה 2 – Building Machine Learning classifiers**