תיאור הפרויקט:

ראשית, יצרנו פונקציה שעברה על כל קבצי הציוצים שקיבלנו והחזירה את 20 המילים הנפוצות ביותר עבור כל אישיות. הסיבה שכתבנו פונקציה זו הייתה לא לשימוש בעת עיבוד המידע, אלא לקבלת תחושה לגבי המילים המאפיינות והביטויים הייחודיים עבור כל אישיות.

עם זאת, ידענו כי עלינו לעבד את המידע כאשר הוא מכיל ציוצים של כל האנשים על מנת לאמן את המסווג כמו שצריך. על כן, בשלב הבא איחדנו את הציוצים מתוך כל קבצי המידע וערבבנו אותו.

לאחר מכן, חילקנו את המידע למידע האימון ומידע הבדיקה – כאשר 80% אקראיים נבחרו בכל פעם להיות מידע האימון והשאר להיות מידע הבדיקה. אותם כמובן חילקנו למידע ולתגיות – כלומר לאישיות שצייצה את הציוץ הרלוונטי.

בשלב זה, ניקינו את המידע על ידי כך שמחקנו את הקישורים (שכן תחילת הקישור נבחרה כמילה נפוצה עבור כמעט כל אחד מהאנשים) וכמו כן גם את כל המספרים בעזרת שימוש בספרייה $textacy$. כמו כן, ניקינו את מילות הקישור והיחס (שסביר שיופיעו אצל כולם במידה שווה) בעזרת הספרייה $sklearn$.

לאחר מכן, השתמשנו ב$TfidfVectorizer$ על מנת לסווג את המידע באמצעות המילים הייחודיות והמאפיינות עבור כל אישיות. קיבלנו מטריצה שמכילה את המילה/ביטוי ואת מספר ההופעות שלהם ומתוך איזה ציוץ הם נלקחו.

בשלב זה יצרנו מסווג מסוג $LogisticRegression$, תחילה מפני שאנחנו יודעות לעבוד איתו ומתוך נוחות. בדקנו מהי השגיאה על ידי ספירת ההפרשים בין התגיות שבy לבין החיזויים של המסווג שלנו חלקי מספר הציוצים. לאחר שקיבלנו שגיאה מסוימת, רצינו לשפר את רמת הביצועים.

לשם כך, בדקנו סוגים שונים של מסווגים – SVM, RandomForest, LDA... כל אחד מהמסווגים הפגין ביצועים פחות טובים מאשר ה- $LogisticRegression$ – או שהשגיאה הייתה גבוהה בהרבה ואף הגיעה קרוב ל100%, או שהריצה לא עצרה.

על כן, החלטנו להישאר עם המסווג מסוג $LogisticRegression$.

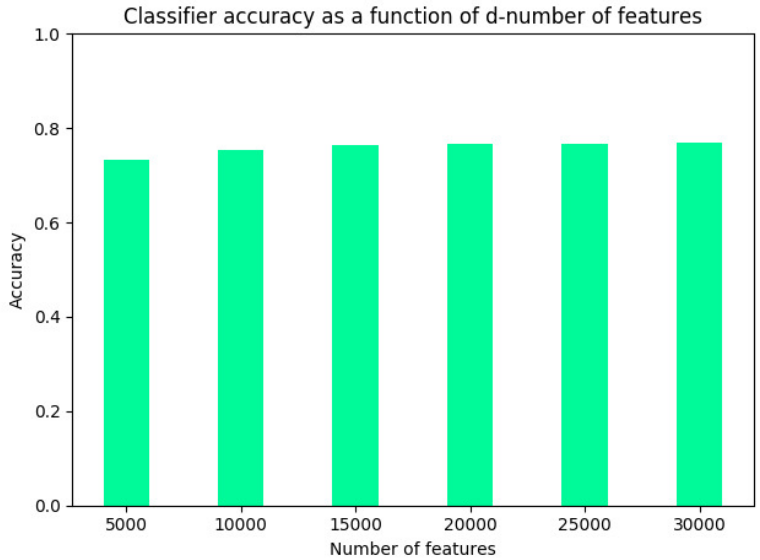
בשלב זה נתקלנו בבעיה מוזרה – אחוז השגיאה שלנו היה כה נמוך שהגיעו לנו פרסים ומענקים עליו! על כן, הבנו שכנראה לא הפרדנו את מידע האימון כמו שצריך ממידע הבדיקה ושהמסווג שלנו משוחד, כלומר שנוצר מצב של overfitting. התמודדנו מספר שעות עם בעיה זו.

לאחר מכן, חיפשנו דרכים נוספות לשפר את הביצועים של המסווג. על כן, רצינו להוסיף מאפיינים נוספים לבדיקה מלבד המילים המאפיינות כל אישיות. שמנו לב כי טראמפ משתמש באותיות גדולות פעמים רבות בציוצים שלו, רונאלדו נוטה לכתוב בשפות נוספות כמו פורטוגזית, איטלקית וספרדית, ליידי גאגא משתמשת באימוג'ים רבים וקים קרדשיאן נוטה לתייג המון ומשתמשת המון בסימן $ (מפני שמשווקת את הקולקציה שלה) ובסימני קריאה רבים.

לשם כך, יצרנו פונקציה שבודקת אילו מהציוצים מקיימים תנאים אלו. רצינו להוסיף אותה למידע שהתקבל מהבדיקה אודות המילים הייחודיות. על כן, היה עלינו להפוך את המידע על המילים למטריצה (על מנת שנוכל להוסיף לה עמודות עם המאפיינים הנ"ל).

לצערנו הרב, בהתחלה אכן הצלחנו להפוך את המידע למטריצה ולאחד אותו עם המידע עבור הערכים הנ"ל, אך בשלב מסוים כנראה שינינו משהו והמסווג לא הצליח לעבד את המידע לאחר שהפכנו אותו למטריצה. לא הצלחנו למצוא את השגיאה שלנו זמן רב ועל כן ויתרנו על בדיקות אלו – כלומר לא הפכנו את המידע למטריצה כך שלא ניתן היה להוסיף אליה עמודות עם מאפיינים נוספים.

בשלב זה, רצינו לבדוק מהו מספר המילים האופטימלי שיש לקחת עבור כל אישיות. לשם כך, יצרנו גרף שבודק מספר שונה של מילים ואת ההשפעה שלהן על אחוז השגיאה בממוצע. הגרף שקיבלנו הוא:



השגיאה האופטימלית התקבלה עבור 2500 מילים ולכן בחרנו ב2500 מילים עבור כל אישיות.

לבסוף, בחרנו לאמן את המסווג ולשמור אותו באמצעות הספרייה $pickle$ על מנת שנוכל להעביר את המסווג כאשר הוא כבר מאומן ומוכן.