**מאגרי מידע**

מאגרי המידע שלנו מורכבים ממשתמשים אמיתיים (כלומר שמאחוריהם עומדים אנשים אמיתיים) ובוטים ברשת החברתית טוויטר. הם מכילים את מזהה המשתמש בטוויטר ואת סיווגו כבוט או לא בוט.

שליפת מידע ואחסונו

כדי לגשת למידע המלא על כל משתמש כמו הציוצים שלו והפרופיל שלו היינו צריכים להשתמש בממשק של טוויטר tweepy . זה הוא api שבהינתן מזהה משתמש בטוויטר נותן גישה מלאה לפרופיל שלו ולציר הזמן שלו ולציוצים שלו. לכל משתמש במאגר ביצענו אחסון של כל המידע שיכולנו לקחת ממנו באובייקטים מסוג user שיצרנו שמאפשרים גישה נוחה ומהירה לכל המידע על המשתמש שנרצה לעשות בו שימוש. לאחר מכן לקחנו את האוביקטים והסיווג שנוצרו לכל המשתמשים ושמרנו בקובץ כדי שלאחר מכן נוכל לטעון את המידע במהירות ובפשטות בזמן ריצת המודל.

סינון מידע

בעת הגישה למשתמשים באמצעות tweepy קיימים משתמשים שהוסרו מטוויטר או שהפכו את הגישה לפרופילים שלהם לפרטיים ולכן לא ניתן לגשת אליהם ולקחת מהם מידע ולכן הסרנו אותם מהמאגר.

בנוסף הניתוח שהמודל שלנו יעשה בהמשך על המידע של המשתמשים דורש ששפת הציוצים שלו תהיה באנגלית ולכן משתמשים שציוציהם לא היו באנגלית הוסרו גם הם מהמאגר.

עיבוד הטוויטים לפני השימוש בword2vec

\*הסבר קצר על word2vec- מדובר ברשת נוירונים רדודה שמקבלת סט רחב של מילים ושולח כל אחת מהמילים לווקטור שונה משלה בעל מימדים גדולים כאשר מילים שחולקות קונטקסט נשלחות לווקטורים שהמרחק ביניהם יותר קטן.

הטוויטים אותם אנחנו מקבלים מהמשתמשים חייבים לעבור "ניקוי" כדי שword2vec יוכל לעבד אותם בצורה טובה ומועילה.

ראשית טוויטים מכילים בדרך כלל תיוג של משתמשים אחרים ברשת על ידי התו @ ולאחר מכן שם המשתמש , לכן כדי שword2vec יוכל לתפוס איזכור של משתמש בתור קונספט במשפט ולא בתור מילה ספציפית ומיוחדת עבור כל אזכור שונה בכל ציוץ החלפנו את אזכורי המשתמשים בהופעת התו @ בלבד.

לאחר מכן שימוש בקישורים בציוצים צריך לעבור אותו תהליך כי אינינו רוצים שword2vec יסתכל על כל קישור כמושג נפרד ושונה בעולם אלא כקונספט הכללי של לשים קישור בציוץ ולכן החלפנו הופעה של כל קישור בתו $ .

לבסוף הסרנו את כל שאר התווים המיוחדים מציוצים שכן הם לא תורמים ואף גורעים מההבנה של word2vec את הקונטקסט של המילים.

לאחר מכן המילים של כל ציוץ עוברות tokenization כלומר הציוץ מפורק לרשימה של המילים במרכיבות אותו כאשר כל מילה או פתגם או ביטוי הופך לאיבר ברשימה זו כלומר לtoken. את הtokens ניתן להכניס לword2vec בתור קלט שהוא יכול להתאמן עליו וגם לפלוט ווקטורים מתאימים בזמן ריצת המודל המלא.