Learning

explained

המשימה

במאה ה- 21 הרשתות החברתיות מהוות נדבך בתהליכים חברתיים המתרחשים בכל העולם; מחקרים מראים שעם הניתוחים המתאימים ניתן להבין ולהשפיע על דעת הציבור בעזרת הרשתות החברתיות.
כך למשל, הרשתות החברתיות היוו חלק ייסודי ביותר בבחירות לנשיאות ארה"ב בשנת 2016 ובתהליך **Brexit** שמבצעת בריטניה בשנים האחרונות.

זה רק טבעי שהאירועים הגדולים הבאים בעולם יעוצבו גם הם על ידי החברתיות ומן הראוי שנמצא כלים מתאימים כדי לנתח זאת. עקב כך, השאלה העולה לעתים בקרב חקר הרשתות החברתיות היא **כיצד תראה מערכת הבחירות לנשיאות ארה"ב של שנת 2020 ?**
ניתן להבין שכדי לענות על שאלה זו, נרצה למצוא בצורה יעילה אילו מבין הטוויטים מדברים על הנשיא **Trump**.
מנגנון יעיל למציאת אותם טוויטים יכול לשמש קבוצות חוקרים למיניהם בכל הנוגע להבנת התגובות ברשת הטוויטר בנוגע להצהרות פומביות של הנשיא **Trump**.

בעקבות כך, החלטנו לנסות להיעזר במידע שניתן לנו מטוויטר על מנת לבנות מודל שידע לחזות אילו טוויטים מדברים על הנשיא **Trump** על פי כמות סימוני ה **favorite** וה **retweets** שטוויט מקבל.
לשם הכנת המודל, כפי שציינו במחברת ה **ETL**, החלטנו לקחת תאריכים שונים בהם **Trump** יצא בהצהרות פומביות הנוגעות להתנהלות ארה"ב בתקופת הקורונה.

האלגוריתם

לשם הכנת המודל שתיארנו, בחרנו להשתמש באלגוריתם הסיווג **Logistic Regression**.
תיאור האלגוריתם על פי וויקיפדיה:

רגרסיה לוגיסטית היא מודל סטטיסטי המתאר קשר אפשרי בין משתנה איכותי/קטגורי, המכונה "המשתנה המוסבר", ובין משתנים אחרים המכונים "משתנים מסבירים". המשתנים המסבירים יכולים להיות איכותיים או כמותיים. המודל מאפשר לאמוד את מידת ההשפעה של שינוי בערכו כל אחד מהמשתנים המסבירים על ערכו של המשתנה המוסבר. במילים אחרות, המודל מאפשר לאמוד מתאמים בין המשתנים המסבירים למשתנה המוסבר. המודל לבדו אינו מספיק כדי לקבוע קשר סיבתי בין המשתנים המסבירים והמשתנה המוסבר.

ניתן להבין כי כאשר אנחנו מתמודדים עם משימת סיווג בינארית, תחת ההנחה כי ניתן לאמוד את הקשר בין המשתנים המסבירים לבין המשתנה המוסבר, האלגוריתם הנ"ל יהווה בחירה מתאימה למשימה.
במקרה שלנו, המשתנים המסבירים הינם (הסבר מצורף על פי **Twitter**):

retweet_count	Int	Number of times this Tweet has been retweeted. Example: `"retweet_count":160`
favorite_count	Integer	*Nullable*. Indicates approximately how many times this Tweet has been liked by Twitter users. Example: `"favorite_count":295`

והמשתנה המוסבר שלנו הוא **Label** המייצג האם ברשימת התיוגים של הטוויט, קיים לפחות תיוג אחד המכיל את המילה **Trump**.

אימון המודל והרצתו

בתהליך אימון המודל תחילה יצרנו עמודת **Label** שמציגה עבור כל טוויט האם ברשימת התיוגים שלו מופיע תיוג המכיל את המילה **Trump** (ערך 0 מציין שהוא אינו מכיל וערך 1 מציין שהוא כן מכיל). כמובן שכדי לקלוט את כל המקרים האפשריים, כל תיוג נבדק בצורת **lower case** כדי לא לפספס מקרים מתאימים.

לבסוף, יצרנו עמודת **features** שמייצגת כוקטור את המשתנים המסבירים שציינו קודם כפי שראינו ב **workshop**. בעזרת עמודה זו ועמודת ה **Label** ניתן להריץ את האלגוריתם **Logistic Regression** מהספריה של **Spark ML**.

את הרצת האלגוריתם ביצענו בשני אופנים שונים –

1. הרצנו את האלגוריתם על כל הדאטא שאספנו בחלק ה **ETL**, ללא התייחסות לעובדה שרוב הטוויטים בדאטא אינם מתאימים ל **Label** 1.
2. ביצענו תהליך **Undersampling** שבוחר בצורה רנדומית רשומות המתאימות ל **Label** 0 (ה **Label** הנפוץ בפער רב בדאטא) כך שכמות הרשומות מ 2 ה **Labels** תהיה שווה. תהליך זה נפוץ בקרב משימות למידה בעולם הדאטא ונדרש בדרך כלל בבניית מודלים על דאטא שאינו מאוזן בין ה **Labels**.

תוצאות המודל ומסקנות – במחברת המצורפת

קישור למחברת ה Learning בDATABRICS

המשך המחקר

בעקבות התוצאות והמסקנות שהצגנו במחברת המצורפת, נרצה לחקור אלגוריתמי למידה נוספים לביצוע אותה מטלה. בנוסף, לשם השגת תוצאות טובות יותר מהמודלים השונים, ננסה לקחת מגוון שונה של **Attributes** שייתכן ומדגישים בצורה טובה יותר האם טוויטים נוגעים אל הנשיא **Trump** או לא.
ניתן להסיק למשל כי אם היו בידינו כלים מתקדמים לעיבוד שפה טבעית, היינו יכולים לנתח בצורה טובה את תכונת ה **text** שנמצאת בטוויטים ובעזרתה להסיק האם טוויטים קשורים אל **Trump** או לא.