חיזוי אישיות על פיType Indicator Myers-Briggs – ערן כץ ואיציק בן שושן

הMyers Briggs Type Indicator היא מערכת סוג אישיות המחלקת את כולם ל-16 סוגי אישיות נפרדים על פני 4 צירים:

הסתגרות (Introversion) (I) – החצנה (Extroversion) (E)

אינטואיציה (Intuition) (N) – חישה (Sensing) (S)

חשיבה (Thinking) (T) – הרגשה (Feeling) (F)

שיפוט (Judging) (J) – תפיסה (Perceiving) (P)

A picture containing text, screenshot, font, parallel

Description automatically generatedלדוגמה, מי שמתאפיין במופנמות, אינטואיציה, חשיבה ותפיסה יסומן כ-INTP במערכת MBTI. דוגמה לכל הסיווגים:

מערכת זו היא אחת ממבחני האישיות הפופולרי ביותר בעולם. היא משמשת בעסקים, באינטרנט, בשביל הכיף, למחקר ועוד.

נרצה לענות על השאלות איך ניתן לסווג אישיות של אנשים לפי פוסטים שהעלו ברשת ובנוסף לחקור מה היחס בין התיוגים השונים.

המאגר:

מאגר המידע מכיל מעל 8600 שורות של מידע, בכל שורה יש שני דברים:

1. הסיווג של בן אדם מסוים. כלומר 4 האותיות שמבחן הMBTI הניב.
2. קטע שמכיל כל אחד מ-50 הדברים האחרונים שהם פרסמו. כל ערך מופרד ב-"|||".

התפלגות הסיווגים במאגר, ניתן לראות רוב משמעותי עבור הסיווגים INFP, INFJ, INTP, INTJ.

A picture containing text, screenshot, plot, line

Description automatically generated

טיפול במידע:

תחילה המרנו את כל הפוסטים לאותיות קטנות.

הפוסטים מכילים הרבה מידע לא נח למודלים למידת מכונה והמרת הטקסט לוקטורים כגון קישורי URL, מספרים, סימנים מיוחדים כמו סימני שאלה, סימני קריאה מקף תחתון ועוד. כל אלו הוחלפו ברווח. לאחר מכן, החלפנו את כל המקרים בהם יש לפחות שני רווחים רצופים לרווח יחיד.

בנוסף, גילינו כי יש הרבה מידע שלא חשוב לסיווג האישיות ואפשר לנקות אותו מהפוסטים. מידע זה הוא מילים הנקראים מילות עצירה או באנגלית stopwords. מילות עצירה הן קבוצה של מילים נפוצות בכל שפה. לדוגמה, באנגלית, "the", "is" ו-"and". מילות עצירה משמשות כדי להוריד מילים לא חשובות ממאגר המידע, מה שמאפשר לנו להתמקד במילים החשובות יותר. לבסוף, כדי שכל המידע יהיה תואם, הפכנו מילים ברבים כמו לדוגמה dogs ליחיד (dog).

המודלים והטכניקות:

לפני שנוכל לאמן את המידע נרצה שהמידע יהיה מתאים לאלגוריתמים עליהם ירוצו. לכן נצטרך להשתמש בLabelEncode() על מנת להתאים מספר לכל סיווג. מכאן 0 עד 15 סיווגים שונים. בנוסף נרצה להשתמש בTfIdfVectorizer זהו אלגוריתם שמשמש להפיכת טקסט לייצוג משמעותי של מספרים המשמש להתאים אלגוריתם מכונה לחיזוי.

לאחר החלוקה לtrain set ולtest set קיבלנו את התוצאות האלו על ארבעת האלגוריתמים הנ״ל:

|  |  |
| --- | --- |
| Accuracy | Model |
| 0.380403 | Multinomial Naïve Bayes |
| 0.542363 | Random Forest |
| 0.653602 | SVM |
| 0.650144 | Logistic Regression |
| 0.393083 | AdaBoost |
| 0.512391 | K Neighbors (82 neighbors) |

A picture containing text, screenshot, diagram, plot

Description automatically generated

ניתן לראות כי התוצאות לא גבוהות במיוחד כאשר הן לא עולות על 65 אחוז במודלים הטובים.

רצינו לשפר זאת.

שיפור התוצאות:

לפי מבחן הMBTI אפשר לראות שיש סה״כ 16 אפשרויות סיווג מכיוון שלכל אות יש שני אפשרויות (2^4) לפי ההגדרה הסתגרות (Introversion) (I) – החצנה (Extroversion) (E),

אינטואיציה (Intuition) (N) – חישה (Sensing) (S) וכו'. גילינו כי אם נחזה כל אות בנפרד ולא נסווג את כל האותיות יחד, נראה שהתוצאות יניבו חיזויים מדויקים בהרבה. לבסוף נאחד את כל האותיות שמצאנו עבור אדם מסויים ונקבל את הסיווג הסופי.

תחילה נראה את היחסים בין האותיות:

A picture containing text, screenshot, rectangle, diagram

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generated

A picture containing text, screenshot, rectangle, diagram

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generated

בהשוואה בין גדלי המדגם בתוך כל אחד מארבעת הצירים, אנו יכולים לראות שלצירי E-I ו-N-S יש גדלי מדגם מאוד לא מאוזנים.

לפני שניגש לעבוד על המידע המחולק, נבדוק קודם מהי התלות בין כל שני אותיות בעזרת מטריצת יחסים

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| J-P | F-T | N-S | E-I |  |
| 90.161- | 0.0695- | 0.0458 | 1.000 | E-I |
| 0.0149 | 0.0809 | 1.000 | 0.0458 | N-S |
| 0.0046 | 1.000 | 0.0809 | 0.06957- | F-T |
| 1.000 | 0.0046 | 0.0149 | 0.1619- | J-P |

נשים לב שאם נסתכל על הערכים המוחלטים של היחסים בין התיוגים, נראה שהם נמוכים (לא עולים על 0.2), מה שמעיד על תלות נמוכה. מכיוון שהתלות נמוכה, נוכל לחזות כל אחד בנפרד ולבסוף לאגד את התוצאות לתוצאה אחת משותפת, מבלי שתהיה פגיעה משמעותית בדיוק התוצאה.

המודלים והטכניקות:

כמו מקודם השתמשנו בLabelEncoder ובTfIdfVectorizer כדי להתאים את המידע למודלים. הTfIdfVectorizer לא השתנה מכיוון שמדובר על אותם פוסטים, אבל הLabelEncoder כן. במקרה הזה, מכיוון שבכל מקום מבין 4 האותיות בסיווג יש שני אפשרויות, E מול I, J-P וכו׳, אז הEncoder יקצה מספר, 0 או 1, לכל שני אפשרויות. אם למשל הסיווג הוא INTP, אז הסיווג במספרים יהיה 1011.

לאחר החלוקה לtrain set ולtest set של כל המידע, לקחנו עבור כל זוג אותיות את החלק שלהם ואימנו אותם על המודלים. קיבלנו את התוצאות הנ״ל:

עבור הTrain:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Total Train accuracy | J-P | F-T | N-S | E-I | Model |
| 0.896 | 0.889 | 0.928 | 0.893 | 0.877 | Logistic Regression |
| 0.762 | 0.607 | 0.812 | 0.863 | 0.765 | Random Forrest |
| 0.920 | 0.897 | 0.924 | 0.944 | 0.918 | SVM |
| 0.801 | 0.724 | 0.874 | 0.863 | 0.766 | Multinomial Naïve Bayes |
| 0.849 | 0.806 | 0.834 | 0.905 | 0.853 | AdaBoost |
| 0.833 | 0.787 | 0.835 | 0.884 | 0.827 | KNeighbors(33 neighbors) |

עבור הTest:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Total Test Accuracy | J-P | F-T | N-S | E-I | Model |
| 0.842 | 0.800 | 0.856 | 0.870 | 0.844 | Logistic Regression |
| 0.742 | 0.613 | 0.715 | 0.858 | 0.782 | Random Forrest |
| 610.8 | 0.806 | 0.853 | 0.919 | 0.869 | SVM |
| 0.772 | 0.652 | 0.798 | 0.858 | 0.782 | Multinomial Naïve Bayes |
| 0.826 | 0.776 | 0.799 | 0.891 | 0.840 | AdaBoost |
| 0.822 | 0.772 | 0.808 | 0.879 | 0.831 | KNeighbors(33 neighbors) |

A graph of different colored bars

Description automatically generated

אפשר לראות שיפור משמעותי בכל המודלים כאשר אימנו על כל זוג אותיות בנפרד, וכמובן שלפי זה יהיה שיפור משמעותי עבור הסיווג הכולל.

המודלים שהניבו את התוצאות הטובות ביותר הם Logistic Regression ו-SVM והתוצאות הגרועות ביותר הוא הMultinomial Naïve Bayes.

זמני הריצה עבור כל מודל:

|  |  |
| --- | --- |
| Time in Seconds | Model |
| 1.3247 | Logistic Regression |
| 26.5856 | Random Forest |
| 439.844 | SVM |
| 0.0523 | Multinomial Naïve Bayes |
| 43.7932 | AdaBoost |
| 1022.1950 | Kneighbors (33 neighbors) |

ניתן לראות שSVM ומודל השכנים הם המודלים שזמני הריצה שלהם ארוכים משמעותית משאר המודלים. מצאנו דרך לשפר את זמני הריצה עבור שני מודלים אלו בעזרת PCA. זמני הריצה המעודכנים עבור שני המודלים האלו עבור 60 רכיבים:

|  |  |
| --- | --- |
| Time in Seconds | Model |
| 10.2771 | SVM |
| 15.8286 | Kneighbors (33 neighbors) |

קובץ main:

הוספנו קובץ main על מנת לבדוק את האלגוריתם שלנו על נתון חדש של מידע. בקובץ הזה אנו מריצים את המודל של Logistic Regression על טקסט חדש שהמשתמש יכול לכתוב. בחרנו להריץ מודל זה מכיוון שתוצאות הדיוק שלו אמנם השניות בגבהם, אך זמן הריצה שלו הכי מהיר. לבסוף מתקבל חיזוי שהוא 4 האותיות שמסמלות את האישיות של המשתמש לפי המבחן של MBTI.