תרגיל 4 בעיבוד שפה טבעית לעברית – סיווג טקסט

אפשר להגיש בזוגות. להגשה עד 17.6.2021.

"חופים הם לפעמים געגועים לנחל" (נתן יונתן)

הסבר

1. התרגיל הנוכחי עוסק בסיווג טקסטים המתארים אטרקציות תיירותיות בישראל.
2. בקובץ israel_attractions.tsv יש מידע על 584 אטרקציות בישראל. לכל אטרקציה מופיעים השדות:
   a) Id – מזהה מספרי
   b) Name – שם האטרקציה
   c) Kind – סוג האטרקציה
   d) Censored_Desc – תיאור מילולי של האטרקציה, ממנו מחקתי חלק מהמילים והחלפתי אותן במילה "מצונזר".
3. עליכם לבנות מסווג שיזהה את ערך Kind רק מתוך ערך השדה Censored_Desc. הסיווג צריך להיעשות בעזרת ספריית sklearn, וספריות נוספות אם אתם זקוקים להן. אפשר להיעזר בחומרים משיעור 6.
4. בשלב ראשון, יש לעשות אקספלורציה של הנתונים: כמה סוגי אטרקציות מופיעים בקובץ? כמה פריטים יש מכל סוג? כמה מילים יש בממוצע ב-Censored_Desc?
5. בשלב הבא, יש להשתמש ב-train_test_split באופן הבא:
   X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
   כדי לקבל קבוצת test שהיא 0.2 מכלל הנתונים.
6. עליכם לעשות preprocessing לטקסט, ולהריץ לפחות שני אלגוריתמי סיווג, שאחד מהם הוא רגרסיה לוגיסטית.
   לכל אלגוריתם יש לחשב את מטריצת הבלבול, precision, recall, F1 על כל תוית ב-test.
7. מהן המחלקות ביניהן התבלבלו האלגוריתמים? מה המחלקה הכי קשה לסיווג? קחו דוגמה שסווגה באופן שגוי ונסו להסביר מדוע.
8. מתוך הרגרסיה הלוגיסטית, הציגו את עשרת המשתנים (מילים או צירופי מילים) החשובים ביותר לסיווג כל מחלקה.
9. איזה אלגוריתם נתן F1 גבוה יותר?

יש להגיש notebook שמכיל את הקוד ותשובות לשאלות (אפשר לענות באנגלית).

בהצלחה,
יובל