למידת מכונה – מטלה תחרותית

הצגת בעיית הלמידה

נתון corpus (מאגר) מתוייג של סיפורים. לכל סיפור יש תגית המציינת את המגדר של כותב הסיפור.

המגדר מיוצג בתור מחרוזת עם 2 אפשרויות:

- 'm' – עבור כותב (male)
- 'f' - עבור כותבת (female)

עליכם לבנות מודל סיווג שבהנתן טקסט של סיפור מסווג את הסיפור לפי המגדר של הכותב.

תאריך הגשת המטלה

את המטלה יש להגיש עד היום האחרון של הסמסטר. כרגע מדובר על ה-1.6, אם הסמסטר ידחה בכמה ימים, כנראה שתקבלו זמן נוסף בהתאם (זמן ההגשה המדויק נמצא במטלה עצמה).

קבצי הגשת המטלה

עליכם להגיש 2 קבצים:

1. competitive_assignment.ipynb - קובץ ה-jupyter notebook המצורף למטלה (ללא שינוי שמו), המכיל את כל הקוד בו השתמשתם לצורך אימון המודל, וסיווג הדוגמאות החדשות.
   - הקוד אמור להריץ את כל השלבים שישמשו לבניית המודל ולסיווג הדוגמאות ב test. עליכם להשתמש בקובץ ה-excel המייצג corpus המתויג לאימון ולסווג את דוגמאות האימון המופיעים ב-corpus הלא מתוייג, כפי שמפורט בפסקה הבאה.
   - שימו לב אתם צריכים לכתוב קוד עובד, מקורי שלכם, שיעבוד גם בסביבה שלנו ויפיק את אותם תוצאות שאתם מצרפים בקובץ ה-csv (המתואר בסעיף הבא).
   - יש ללוות את הקוד שלכם בהערות הסבר בגוף הקוד.
2. classification_results.csv - קובץ csv המכיל את הסיווגים שלכם עבור כל דו' ב-test. הקובץ צריך להכיל 2 עמודות:
   - 'test_example_id' – מזהה המסמן את דוגמת ה-test (לפי הסדר המקורי)
   - 'predicted_category' – עמודה המכילה את סיווג דוגמאות ה-test. התאים בעמודה, יכולים להכיל שני ערכים אפשריים (המייצגים את שני הסיווגים האפשריים):
     - 'm' – עבור כותב
     - 'f' - עבור כותבת

הסברים על ניקוד המטלה

סך הנקודות שניתן לצבור: 27 נקודות.

על מה אתם נמדדים?

עיקר הדגש בניקוד המטלה ינתן על איכות המודל אותו אתם בונים עבור מודל סיווג המגדר.

כדי להעריך את איכות המודל אותו אתם בונים יחושב כשהשוואה בין התוצאות שאתם מגישים בקובץ *classification_results.csv* לבין התוצאות המצופות (שלא חשופים לסטודנטים).

המדד הנבחר להערכת איכות המודל הוא מדד f1.

תזכורת מדד f1

f1=2*(precision * recall)/(precision + recall)

המדד שנבחר להעריך את איכות המודל - Average-f1

f1_male – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבים כמחלקה החיובית ואת הכותבות כמחלקה השלילית.

f1_female – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבות כמחלקה החיובית ואת הכותבים כמחלקה השלילית.

<u>Average_f1</u> – יחושב כך: **Average_f1=(f1_male + f1_female)/2**

אז כיצד נשתמש ב- Average-f1 וכיצד ננקד את המטלות?

- ע"מ לקבל 7 נקודות – יש להגיש את הקבצים: את קובץ התוצאות *classification_results.csv* ***ואת קובץ*** מחברת המטלה עם הקוד שלכם - competitive_assignment.ipynb, כאשר יש קוד עובד, אישי עם הערות המסבירות אותו וכאשר התוצאות ב-*classification_results.csv* תואמות את אלה שנריץ בקוד שלכם ב-competitive_assignment.ipynb
- כדי לקבל יותר מ-7 נקודות עליכם להשיג Average_f1 של לפחות 0.425
- Average-f1 של 0.425 יזכה אתכם ב-9 נקודות על המטלה (תוצאה זו צפויה להתקבל עם מאמץ מינימלי ללא שום ניסיון לשפר)
- Average-f1 של 0.7 יזכה אתכם ב20 נקודות
- כל תוספת של 0.025 עבור תוצאת ה- Average-f1 (מעבר ל- 0.425 ועד 0.7) תזכה בנקודה נוספת (כך למשל של 0.45 תזכה ב-10 נקודות, תוצאה של 0.5 תזכה ב12 נקודות ותוצאה של 0.575 תזכה ב-15 נקודות).
- כל נקודה מעבר ל20 נקודות תקבע לפי דירוג תחרותי. כאשר הצפי של אלו שיגיעו לרף התחרותי הוא לגרף המתפלג נורמלי (כלומר אם כולם יקבלו תוצאה מקסימלית אף אחד לא יוכל לקבל ניקוד תחרותי).

הקבצים המצורפים לרכיב המטלה

המטלה מלווה ב3 קבצים:

קובץ 1. Corpus מתויג – עבור ה-training

שם הקובץ: annotated_corpus_for_train.xlsx

קובץ excel של ה-corpus המתויג. מדובר בקובץ שמכיל train data, בצורה גולמית ושיש להפוך אותו ל- feature vectors כפי שלמדנו. הקובץ מכיל 2 עמודות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'gender' – עמודה המכילה את המגדר של כותב/ת הפסקה

<u>השימוש בקובץ ה-corpus המתויג עבור ה-training</u>- השתמשו בקובץ על מנת לבנות מודל סיווג שתפקידו לסווג את המגדר של פסקה סיפורית כנ"ל.

קובץ 2. Corpus לא מתויג -עבור סיווג ה-test

שם הקובץ: corpus_for_test.xlsx

קובץ excel נוסף, המכיל פסקאות סיפוריות. הקובץ מכיל דוגמאות חדשות אותם יש לסווג.

הוא מכיל את העמודות הבאות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'test_example_id' –המסמן את דוגמת ה-test

<u>השימוש בקובץ ה-corpus **הלא** מתויג עבור סיווג ה-testest</u> - לאחר בניית המודל (מהדוגמאות ב-corpus המתויג שמוסבר בסעיף הקודם), עליכם להשתמש באותו תהליך של preprocessing (שבצעתם על ה-training) ובאותו vectorization על הסיפורים שבמסמך ה- corpus הלא מתויג (הקובץ שתואר בסעיף זה), ולהשתמש במודל אותו אימנתם על מנת לסווג את המגדר של כותב/ת הסיפור.

את הפלט יש לשמור בקובץ ה-csv כפי שמתואר לעיל.

קובץ 3. מחברת הגשה ריקה להגשת התרגיל

שם הקובץ: competitive_assignment.ipynb

המחברת שתריצו את הקוד (כפי שמתואר לעיל בקבצי ההגשה לעיל). במחברת אינה מכילה כל קוד מחייב, מלבד המלצות לטעינה וכתיבת הפלט (אותו יש להגיש גם כן כקובץ נפרד). אתם יכולים להשתמש בכל קוד שתרצו (כולל ה-tokenizer המופיע בכלים בעברית במודל).

הסברים נוספים

יתווספו, אם יהיה בכך צורך

בהצלחה לכולם :-)