עבודת בית #3 - Text Classification

תומר רמון 204621510
דניאל נחשוני 303111744
נועם אפרת 307903054

קישור לגיט: https://github.com/danielnachshoni/TextClassifier

היכולת לסיווג טקסט הינה משמעותית ובעלת ערך רב לרשתות חברתיות גופים ממשלתים ועוד.
באלגורתים שפותח ע"י Fernando López, אנו רואים אפשר לבצע סיווג טקסט בעזרת למידה עמוקה.

שיטת העבודה של אלגוריתם זה מחולקת לשני חלקים :
הראשון הוא עיבוד הדאטה , הנתונים נלקחו מ kaggle contest, דאטה-סט המכיל ציוצים (Twitter) לגבי אסונות שהתרחשו או לא התרחשו.

חלק א' - עיבוד הדאט
א. הפרדת משפטים למילים בודדות.
ב. השמת אינדקס מספרי עבור כל מילה במשפט . נציין בקצרה כמה מהפרמטרים אשר קובעים את הערך המספרי של מילים המופרדות.
max_len = כמה תווים יכולים להימצא במילה בודדות.
max_word = מתייחס למילים המופיעות בתדירות הכי גבוהה.

חלק ב'- אימון המודל:
האינדקסים של הדאטה-סט משמשים כקלט למודל.
המודל בעצמו "מפרק" את הקלט לתתי פרמטרים כמו גודל אוצר המילים , מימדי הוקטור עבור הפלט ומימד וקטורים שאינם עומדים ברף .
לאחר מכן מתחיל שלב האימון , מספר החזרות באימון מוגדר כמספר ה"אפוקות"(epochs) , מוגדרת פונקצית הפסד, ערכי סף .

לסיכום במאמר זה הוסברו ומומשו הגישות השונות של סיווג טקסט באמצעות למידה עמוקה, שילוב של יתרונות pytorch ושיטת tokenization מתקדמת קידמו את איכות המודל ועזרו לו להפיק תוצאות טובות מהצפוי. עם זאת ע"פ מחקרים אחרונים ניתן אפילו לשפר עוד יותר את המודל הזה באמצעות שדרוג הארכיטקטורה ל (Transformer) במקום שיטת הtokens.

<u>שלבי מימוש האלגוריתם:</u>

1,תחילה עשינו יבוא של הספריות הדרושות TORCH התקנו את TORSCHDATA על מנת לעבוד עם DATASETS

2. TORCH עובר על מערך מילים וממיר בעזרת TOKENIZER כל מילה לערך מספרי מתאים בהתבסס על טבלת מידע שהוגדרה באוצר המילים.

3.על ידי שימוש ב PYTORCH יוצרים איטרטור בעזרת DATA LOADER בעזרתו נוכל לעבור לעבור בכל איטרציה על קבוצה מצומצמת של נתונים ולעבד אותם.

4. הגדרת המודל: שכבת המודל בנויה משכבת הטמעה שמחשבת את הממוצע של ערכי כלל ההטמעות, בנוסף המודל מכיל שכבות לינאריות שמטרתם לסווג את המידע והנתונים של ההטמעות.

5. הגדרת ערכת נתונים, מגדירים את התווית לפיה נרצה לסווג את הטקטס ומספר ההמחלקות יהיה כמספר התויות.

את המודל בונים עם מימד 64 וגודל אוצר המילים שווה לאורך המופע של אוצר המילים ומספר המחלקות שווה למספר התוויות שערכת הנתונים

6.הגדרת פונקציה לאימון שעוברת על הנתונים ומאמנת את המודל שלנו

הגדרת פונקציה שלוקחת את המודל המאומן ומעריכה את טיב האימון של הערכת המודל.

7. מבצעים חלוקה של ה-DATASET המיועד לאימון, כאשר 95% מאותו DATASET יופנה אל עבר אימון המודל ו-5% אל עבר מנגנון אימות. בנוסף, מתבצע שימוש ב CrossEntropyLoss שזהו קריטריון עבודה שמאחד שני משתנים לידי Class אחד והוא מיעל בשיטת ירידה סטוכסטית הדרגתית .
עקומת הלמידה מוגדרת להיות 5.0 אשר מותאם לקצב למידה לאורך תקופות ממושכות.

8.הערכת המודל עבור DATASET - מבצע בדיקות של תוצאות אימון המודל מול ה DATASET שהוכנס ומדפיס את תוצאות הבדיקה.

9. הכנס קלט חדש - מכניסים טקסט אקראי שאינו נמצא ב DATASET שאיתו אימנו את המערכת.
עבור המערכת נבחר המודל שעבורו קיבלנו את תוצאות הבדיקות המדויקות ביותר ובעזרתו בודקים את הטקסט החדש שהוכנס.

<u>תוצאות האלגוריתם:</u>

הרצנו את האלגוריתם על שני טקסטים מקטגוריות שונות. טקסט אחד נלקח מאתר החדשות CNN ואכן האלגוריתם הגיע למסקנה נכונה בדבר הטקסט. ("Australian voters have delivered a sharp rebuke to the center-right government, ending nine years of conservative rule, in favor of the center-left opposition that promised stronger action on climate change.")

בנוסף, ניסינו להריץ על טקסט עצמאי 'Ronaldo has scored.', והאלגוריתם גם במקרה זה הגיע למסקנה הנכונה.