עיבוד שפה טבעית – פרויקט

מטרת הפרויקט

הקניית כלים פרקטיים לצורך יישום תהליכי למידת מכונה על מטלות עיבוד שפה טבעית.
תחום עיבוד השפה הטבעית מתפתח בקצב מהיר ונשען על כלים חזקים מאוד, מודלים מתקדמים ושיטות חדשניות שבהם נעשים שימושים בחברות הגדולות והמובילות במשק כדוגמת Microsoft, Google, OpenAI ועוד.

Dataset

ה-dataset שאיתו תעבדו מכיל ביקורות על סרטים מ-IMDB עם ה-sentiment המתאים עבור כל ביקורת. תוכלו להיעזר בקובץ ה-README על מנת להבין יותר על ה-dataset.

חלק א' (30%)

בחלק זה תיישמו רשת נוירונים מסוג LSTM ו-GRU על מנת לבצע משימת Sentiment Analysis ותדרשו להצבת מערכת יציבה וגנרית שתאפשר אימון של מגוון מודלים עם שינוי פרמטרים שונים לצורך השוואה.

עבור כל רשת יש ליישם שימוש באחת משיטות ה-embeddings שלמדתם בקורס.

בנוסף, עבור כל רשת יש לבחון את ההבדלים (אם קיימים) עבור שינויי hyperparameters שונים ולהציג את התוצאות השונות.

כמו כן יש לנתח ולעבד את המידע הקיים בכל דרך העולה על דעתכם ולהסביר את ההיגיון לכך.

שימו לב, יש לבצע את חלק זה של הפרויקט באמצעות ספריית PyTorch.

סעיף בונוס: צרו Embeddings משלכם על הנתונים וערכו השוואה אל מול הקיימים. הציגו את תוצאות ה-Embeddings שלכם והסבירו.

בחלק זה תימדדו על:

1. Functionality – על הקוד למלא את דרישות המטלה באופן מלא.
2. Readability – על הקוד להיות קריא וברור.
3. Quality – על הקוד להיות מאורגן, מודולרי וכזה המאפשר לשחזר את תוצאותיו באופן פשוט. כמו כן, על הקוד לקבל כקלט פרמטרים בפורמט שיופיע עם הנתונים של המטלה.

4. Comparison and Research – השוואה בין ארכיטקטורות שונות של הרשת, שינויי hyperparameters והשוואה מול validation set שעליכם להגדיר. בחלק זה לא תימדדו על תוצאות המודל אלא על תחקורן.