

קורס NLP - שיעורי בית 2

תאריך הגשה: 30/6 בשעה 23:59

המשימה

לאמן מודל שמסווג ביקורות על סרטים לשתי קבוצות: ביקורות חיוביות וביקורות שליליות.

הדאטא

מאגר של 50,000 ביקורות מאתר IMDb. הביקורות החיוביות אלו ביקורות שקיבלו ציון שגבוה מ-7, והביקורת השליליות אלו ביקורות שקיבלו ציון שנמוך מ-4. לתיקיית תרגיל הבית מצורפים שני קבצים: `IMDB_train` ו-`IMDB_test`. השתמשו בנתונים שבקובץ `IMDB_train` לבניית המודל ולאימונו, ובקובץ `IMDB_test` לבדיקת הביצועים של המודל.

הכלים

אתם נדרשים להשתמש ב-Word Embeddings שאומנו מראש. יש לכם את החופש להחליט איזה באיזה סוג של Word Embeddings תרצו להשתמש. לבניית המודל, אתם נדרשים לאמן רשת נוירונים. יש להגדיר בקוד `seed=23` עבור הספריות `numpy` ו-`torch`.

התוצר

יש להגיש דו"ח מפורט שמסביר את התהליך שעשיתם בבניית המודל. קובץ הדו"ח יקרא `report_id1_id2.pdf`. יש לפרט מדוע בחרתם להשתמש ב-Word Embeddings הספציפיים בהם השתמשתם וכיצד בניתם את רשת הנוירונים. יש לפרט איזה פרמטרים ואיזה היפר פרמטרים בחרתם למודל. יש לצרף גרפים המתארים את פונקציית ה-Loss לאורך זמן האימון, הן עבור סט האימון והן עבור סט המבחן. יש לצרף גרפים המתארים את אחוזי הדיוק של המודל שלכם לאורך זמן. יש לדווח מה היה אחוז הדיוק אליו הגעתם בתום אימון המודל.

יש להגיש תיקיית קוד בשם `code_id1_id2.zip` המכילה את קבצי הקוד שלכם. אנא הקפידו על [כתיבה מסודרת וברורה](#) של הקוד הכוללת הערות, כדי שנוכל לספק משוב איכותי על הקוד שלכם.

הציון

65% דו"ח
15% איכות הקוד
20% ביצועים

שאלות

- בפורום השאלות שבמודל. אני אקבל מייל על כל שאלה שתשלח בפורום, ואשתדל לענות עליהן בהקדם.
- שעת קבלה: יום ראשון בין 12:00 ל-16:00, בקישור הקבוע שבמודל.