עיבוד שפה טבעית – פרויקט סיום

מטלת הקורס תהיה מטלה יישומית של החומר והפלטפורמות הנלמדות לאורך הקורס. ציון הקורס יהיה מבוסס על הצגת הפרויקט והתוצאות שלו בשילוב ציון על הפרויקט. ציון הפרויקט ישוקלל יחד עם מטלת הRL יחד עם הגנה על הפרויקט.

ציון הקורס יחושב כך:

$$Final\ Grade = 0.7 * \overbrace{(0.3 * Part1 + 0.7 * Part2)}^{NLP} + 0.2 * RL + 0.1 * \text{הגנה}$$

מטרת הפרויקט

מטרת הפרויקט היא להקנות כלים פרקטיים ככל האפשר לצורך יישום תהליכי למידת מכונה על מטלות עיבוד שפה טבעית. עולם עיבוד השפה הטבעית מתפתח בקצב מהיר מאוד ונשען על כלים חזקים מאוד שבהם נעשה שימוש בכלל החברות הגדולות – Google, OpenAI, Facebook ועוד. נציין גם כי המון כלים נפוצים פותחו ע״י חברות אלה ולכן הכרת הכלים הללו חשובה ביותר לצורך התברגות בתעשייה.

מבנה הפרויקט

הפרויקט יחולק לשני חלקים עיקריים כאשר שניהם יתמקדו בסט הנתונים שיצורף – 37 וחצי אלף ביקורות חיוביות ושליליות כאשר להן המטלה תהיה ביצוע סיווג לביקורות חיוביות ושליליות.

חלק א (30%)

חלק א של הפרויקט יכלול יישום רשת נוירונים מסוג LSTM / GRU לבחירתכם לצורך מטלת הסיווג.

מטרות חלק זה של הפרויקט הם הצבת מערכת יציבה וגנרית שתאפשר אימון מגוון מודלים עם שינוי פרמטרים שונים לצורך השוואה.

בחלק זה תימדדו על כמה פרמטרים:

1) גנריות הקוד – הקוד אמור לקבל כקלט פרמטרים בפורמט שיופיע עם הנתונים של המטלה.
2) איכות הקוד – הקוד צריך להיות קריא, מודלורי וכזה שהייתם רוצים להראות למעסיק אפשרי בעבודה. את התוצאות נרצה גם לשחזר בקלות ולכן גם אספקט זה יבחן.
3) השוואה ומחקר - השוואה בין ארכיטקטורות שונות של הרשת, שינויים בהיפר פרמטרים שונים והשוואה לסט ולידציה שעליכם להגדיר. נדגיש כי בחלק זה לא תימדדו על טיב התוצאות, אלא על תחקורן.

שימו לב, חלק זה של הפרויקט מיועד לצורך הבנת התשתית הבסיסית של PyTorch, אנחנו כן רוצים לראות השקעה בחלק זה של הפרויקט, אך חשוב להבין כי עיקר העבודה אמור להעשות בחלק השני של הפרויקט.

תצטרכו לבחון את אפקטיביות צורות ה embedding השונות שלמדתם בקורס. תצטרכו ליישם 2 שיטות שונות לembedding ולהציג את התוצאות השונות עבור כל שיטה.

חלק זה יועלה למודל כקישור לגיט שבו יהיה קובץ הסבר קצר.

סעיף בונוס

צרו Embeddings משלכם על הנתונים וערכו השוואה אל מול הקיימים. הציגו את התוצאות של הEmbeddings שלכם והאם הם הגיוניים.

חלק ב (70%)

חלק ב של הפרויקט יהיה העיקרי וישלב כמה אלמנטים חשובים שלמדתם בקורס שיהיה עליכם ליישם הלכה למעשה.

בחלק זה של הפרויקט תצטרכו ליישם אימון של מודלים מאומנים, שלמדתם או חקרתם באינטרנט, לצורך המטלה הספציפית שלנו. תצטרכו לקחת כ2 מודלים לפחות ולהשוות את התוצאות שלהם.

בנוסף על ההשואה הפשוטה תצטרכו לבצע כיוון למודל בלפחות 2 דרכים ולערוך השוואה בין הדרכים.

את המודלים תוכלו לקחת מהספרייה Hugginface וליישם אותם בעזרת PyTorch – lightning או Pytoch.

את חלק זה של הפרויקט תצטרכו לכתוב במבנה מאמר אקדמי בפורמט המאמרים שיעלו למודל. המאמר לא יהיה מעבר ל6 עמודים.

מטרות חלק זה של הפרויקט זה להשתמש בשיטות הפופולריות היום לפתרון בעיות בעולם עיבוד השפה הטבעית. בנוסף על כך, עולם הML מבוסס מחקר ולכן אנו מאמינים מאוד בקריאת מאמרים והבנתם לצורך יישומם.

בחלק זה תימדדו על הפרמטרים הבאים:

1) איכות הקוד
2) השוואה ומחקר
3) שימוש במדדים מתאימים לבעיה

כתיבת המאמר – מבנה, כתיבה בעברית רהוטה, והסברים על חלקי הפרויקט, בחירותיכם והניסויים שערכתם. מה שלא כתוב במאמר לא יחשב בציון הסופי.

חלק ב יוגש למודל כZIP עם הקוד שלכם, עם המשקולות המאומנות של המודל (הקוד יקבל וירוץ עם המשקולות המאומנות) ועם המאמר שכתבתם בפורמט PDF.

<u>ניקוד</u>

<u>חלק א</u>

- 10% גנריות הקוד
- 20% איכות הקוד
- 30% ניתוח התוצאות
- 40% העמקה וחשיבה
- עד 15% בנוסף עבור פתרון הבונוס

<u>חלק ב</u>

- 20% איכות הקוד
- 20% ניתוח התוצאות
- 50% העמקה וחשיבה
- 10% כתיבת המאמר
- עד 10% בנוסף עבור תוצאות טובות

<u>תאריכי הגשה לחלקי הפרויקט</u>

חלק א – 7.4.22

חלק ב – סוף הסמסטר 10.6