עיבוד שפה טבעית – פרויקט 1

מטלת הקורס תהיה מטלה יישומית של החומר והפלטפורמות הנלמדות לאורך הקורס.

פרויקט הקורס יוגש בזוגות בלבד (!!!!!!!!!!!!!!!!!!!!!!!!)
הגשה בקבוצות של יותר/ פחות מ2 בקבוצה ללא אישור מפורש בשל נסיבות חריגות תגרור הורדה משמעותית בציון.

מטרת הפרויקט

מטרת הפרויקט היא להקנות כלים פרקטיים ככל האפשר לצורך יישום תהליכים בסיסיים בלמידת 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, Meta ועוד..

נקודה שחשוב להדגיש – בשונה מלפני שנתיים, אז עיקר הפרויקט היה לכתוב את הקוד ולהבין איך לתרגם לקוד את התיאוריה בהרצאות, המצב השנה אחר (LLMs). לכן גם הדגש בבדיקה יהיה אחר. שימו לב לנקודות בהוראות שמתייחסות למתן הסברים, פירוט על תהליכי חשיבה, ניסויים שביצעתם (ועבדו/ לא עבדו). זה מהותי.
פרויקט מושלם מבחינת קוד ותוצאות אבל בלי הסברים, שיקולים, פירוט וכו' וכו' יקבל ציון חלקי (ולא טוב).

מבנה הפרויקט

הפרויקט יתמקד בdataset הבא:

https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification/data

הנתונים כוללים כ50,000 רשומות מתויגות – עליכם לבצע סיווג על גבי הנתונים.

חלק א (30%)

חלק א של הפרויקט יכלול ניתוח של המידע הקיים ועיבוד שלו.

תבצעו ויזואליזציה רלוונטית, ניתוח סטטיסטי רלוונטי וכל העולה על דעתכם (EDA). תהיו יצירתיים ויסודיים.

לאחר הניתוח והסקת מסקנות כאלה ואחרות, עליכם לעבד את הטקסט ולערוך אותו בהתאם לעולם הבעיה שניתן לכם. עשו איזה עיבודים העולים על רוחכם, כל עוד קיים הגיון מאחוריהם.

פרטו ושימו דגש על כיוון המחשבה שלכם, נקודות שבדקתם.ן והצליחו (ובעיקר אלו שלא הצליחו!), מה עמד מאחורי המחשבה שהניעה אתכם.ן וכו' וכו'.

חלק זה יהיה מצורף לאותו GIT repo שבו יהיה הקוד של חלק ב'.

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 (fine tuning, transfer learning) של מודלים מאומנים, שלמדתם או חקרתם באינטרנט, לצורך המטלה הספציפית שלנו.

תצטרכו לקחת **כ2 מודלים לפחות** ולהשוות את התוצאות שלהם.

את המודלים תוכלו לקחת מהספרייה Hugging face וליישם אותם בעזרת – PyTorch כפי שראינו בתרגולים..

את שני המודלים שבחרתם תצטרכו לאמן בשני אופנים:

1. בצעו Fine-tuning באמצעות "הקוד המלא" שראינו בתרגול 4.
2. בצעו Fine-tuning באמצעות הספריות של HF כפי שראינו בתרגול 5.

עבור שני הסעיפים ושני המודלים אתם נדרשים לבצע HP tuning באמצעות Optuna ולתעד את הניסויים שלכם באמצעות W&B בצורה מפורטת ומוסברת כפי שראינו בכיתה. שימו לב שאופן ההתעמקות שלכם בכלים של HF ואופן השימוש בספריות ובפונקציות השונות יבחנו ויינתן עבור כך משקל בציון.

בדומה לתוצאות שראינו בכיתה, אתם תקבלו תוצאות גבוהות מאוד מההתחלה. אני אחפש לראות את תהליך המחשבה שלכם ומה עשיתם על מנת לשפר עוד את התוצאות. אל תסתפקו בתוצאות ראשוניות.

בנוסף על ההשוואה הפשוטה תצטרכו לבצע כיוון לכל מודל **בלפחות 3 דרכים** ולערוך השוואה בין הדרכים (פרמטרי ההשוואה נתונים לבחירתכם ולהבנתכם).

את חלק זה של הפרויקט תצטרכו לכתוב במבנה מאמר אקדמי (מומלץ באמצעות תוכנת Overleaf – תתייעצו עם ה LLM הקרוב לליבכם. או ב WORD כמובן למרות שאני לא ממליץ) בפורמט המאמר לדוגמא שעלה למודל.
המאמר לא יהיה מעבר ל6 עמודים.

כתיבת המאמר – מבנה, הסברים על חלקי הפרויקט, פירוט שיקולים, בחירותיכם (!) והניסויים שערכתם.
מה שלא כתוב במאמר לא יחשב בציון הסופי.

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.

חלק ב יוגש למודל כ PDF שיכלול את המאמר עם קישור לGIT שיכיל את הקוד שלכם.
ב GIT יהיה בניתוב הנכון את הצ'קפוינט עם המשקולות המאומנות של המודל (הקוד יקבל וירוץ עם המשקולות המאומנות).
אני הולך לעשות clone ולהריץ את הקוד לפי ההוראות שלכם שיהיו ב readme. תוודאו שהכל תקין ורץ ללא בעיות.
אם יש בעיה עם העלאת הצ'קפוינט ל GIT יש להעלות אותו למודל יחד עם ה PDF או לשלוח במייל עם הוראות הרצה מסודרות. קוד לא תקין (=ללא דוקומנטציה מסודרת, בעיות הרצה וכו' וכו') לא ייבדק.

<u>ניקוד</u>

- 20% איכות הקוד = 5% גנריות + 5% ארכיטקטורת קוד נכונה, 10% קוד רץ באופן חלק וללא בעיות.
- 20% חלק א' = 10% ניתוח מקדים, גרפים, מסקנות ממה שמצאתם.ן וכו'.
- 15% = העמקה וחשיבה, ייחודיות, פירוט תהליך העבודה ונקודות מעניינות, שיקולים מרכזיים, נימוקים.
- 25% = מחקר וניסויים, תיעוד ב W&B, הצגת Reports מסודרים, מסקנות.
- 10% = כתיבת המאמר.
- 10% = תוכן מעבר לדרישות היבשות (בהצלחה 😊).

<u>תאריכי הגשה לחלקי הפרויקט</u>
תאריך הגשה לכל הפרויקט – 5.8.

תאריך הגשה לחלק ב' יתפרסם בהמשך יחד עם המטלה.

תאריך להגנה על הפרויקט ייקבע בהמשך.