**בינה מלאכותית**

**טוקניזציה**

מודלי שפה אינם מיצגים טקסט כרצף אותיות (מחרוזת) אלא כרצף של מילים, חלקי מילים או צירוף אותיות בעל משמעות (למשל הסיומת ים בעברית).

כשיוצרים טוקנייזר, הוא מוצא את רצפי האותיות הכי נפוצים בסט האימון ונותן לכל אחד מהם מזהה (טוקן) בצורת מספר שלם ואי שלילי.

לטוקנייזר שתי פונקציות מרכזיות:

Encode: הטוקנייזר מקבל מחרוזת ומחזיר רצף של מזהיי טוקן לפי הסדר בהם הם מופיעים בטקסט.

Decode: הטוקנייזר מקבל רצף של מזהיי טוקן ומחזיר ומתרגם אותם לטקסט.

כל מודל שקולט ו\או מייצר שפה מאומן בהינתן טוקנייזר – הטוקנייזר מוגדר לפני תחילת אימון המודל ולא משתנה אף פעם.

**אימון ראשוני לניתוח שפה בעזרת יצירת שפה**

**Generative Pre-Training (GPT) for Natural Language Proccessing**

בהינתן רצף של טוקנים, המודל חוזה כל טוקן בהתבסס על הטוקנים הקודמים בטקסט.

המודל אומנם מקבל את כל הטוקנים אבל בעזרת מסכת הסתכלות קדימה (look ahead mask), כל טוקן תלוי בטוקנים שבאו לפניו בטקסט בלבד.

המודל לומד לחזות איזה טוקן נמצא בכל מקום (מתוך רשימה ידועה של טוקנים).

כאשר:

i המזהה של הטוקן שנמצא במקום ה i.

היא ההסתברות של הטוקן במקום ה t+1 להיות הטוקן שהמזהה שלו i.

seq len הוא אורך הרצף שהמודל מקבל.

המודל מצמם את ה categorical entropy הממוצע לרצף שאותו הוא מנסה לחזות.

שימו לב שהמודל לא מתייחס לטוקן הראשון, זאת מכיוון שאין אפשרות לחזות את הטוקן הראשון בהתבסס על הטוקנים הקודמים כי אין טוקנים שמגיעים לפני הטוקן הראשון.

משימה זו היא משימת לימוד בהנחיה עצמית (**Self Supervised Learning**):

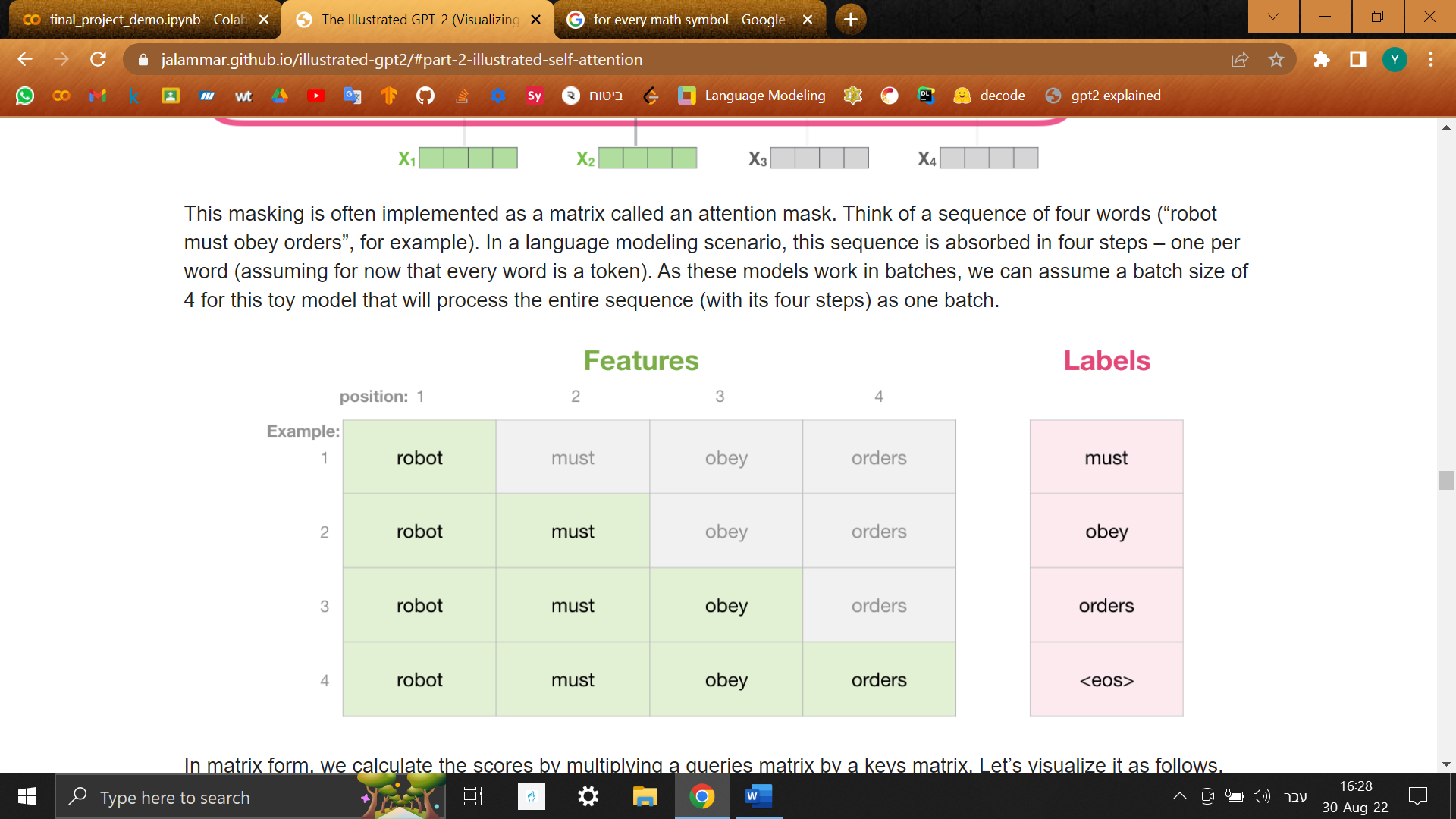
הנתונים עצמם אינם מחולקים לצמדים של קלט ופלט והמודל לומד לחזות או ליצור חלק אחד מהדוגמה בהינן חלק אחד מהדוגמה לכל דוגמה בסט הנתונים.

זאת בניגוד לשיטות למידה פופולריות אחרות:

לימוד מונחה (**Supervised Learning**) בו לכל דוגמה בסט הנתונים יש צמד קלט ופלט. לדוגמה מודל שמסווג תמונות שיש בהן מכוניות לתמונות שאין בהן מכוניות.

לימוד לא מונחה (**Unsupervised Learning**) בו אין חלוקה לקלט ופלט בכלל. לדוגמה: אם בסט הנתונים כל דוגמה היא ווקטור אז אפשר לאמן מודל שמוצא קבוצות של ווקטורים קרובים ומשייך אותם לאותה מחלקה.

חלקוה של המשפט "robots must obey orders" :



**למידה רב שלבית – Transfer Learning:**

המשימה שלחיזוי הטוקן הבא בטקסט אינה חשובה בפני עצמה. האימון למשימת מידול שפה טבעית הוא אימון ראשוני שאחריו מגיע אימון למשימה ספציפית (למשל: סיכום טקסט, תרגום בין שפות, מענה בשירות לקוחות ועוד משימות רבות).

הרעיון מאחורי אימון כללי שלאחריו אימון למשימה ספציפית (down stream task) הוא שהמודל לומד להבין שפה כללית וליצג טקסטים באופן כללי – מה שיעזור מאוד באימון למשימות ספציפיות עם פחות נתונים ופחות כוח חישוב.

גישה זו היא סטנדרטית באקדמיה ובתעשייה בשנים האחרונות.

האימון החוזר על משימה ספציפית נעשה כמו האימון הכללי – בלי התייחסות לקלט ופלט וזאת על מנת שהמודל ילמד לעבד את המידע שמופיע בקלט.

שימוש במודל ליצירת טקסט:

בהינן טקסט שנקלט מהמשתמש:

תחילה המודל מתורגם לרצף של טוקנים על ידי הטוקנייזר.

המודל תחילה מקבל את הקלט ולאחריו טוקן רנדומלי.

המודל חוזה את הטוקן שנמצא במקום האחרון ברצף בהתבסס על הטוקנים שלפניו ולכן זאת התחזית של המודל לטוקן הראשון בפלט.

הטוקן הבא נדגם (ראה חלק על שיטות דגימה) מההסתברות שהמודל יצר.

לאחר מכן המודל מקבל את הקלט ואת הטוקן הנגדם ומוצא את הטוקן הבא וכך הלאה.

מספר הטוקנים שנוצרים יכול להיקבע בכמה דרכים:

1. על ידי המשתמש.
2. מספר הטוקנים הנוצרים יהיה שווה למספר הטוקנים בקלט.
3. על ידי מודל שמקבל קלט וחוזה את אורך הפלט.
4. מתוך התפלגות של הטוקנים הפלטים בסט הנתונים.

**One Hot representetion:**

יצוג one hot ממיר אינדקס של מחלקה (במקרה שלנו טוקן) שהוא מספר שלם ואי שלילי (אפס או חיובי) לווקטור בינארי שאורכו מספר המחלקות בנתונים (במקרה שלנו מספר הטוקנים שהטוקנייזר שומר במילון).

יצוג one hot הוא בעצם ווקטור של הסתברות של קטגוריה ידועה.

היצוג נוצר לפי הפסודו-קוד:

one\_hot(class\_id, numer\_of\_cllasses):

ans = [0 for \_ in range(numer\_of\_cllasses)]

ans[class\_id] = 1

return ans

דוגמה:

אם המילון של הטוקנייזר הוא {אני: 0, אוהב: 1, גלידה: 2} אז המחזרות "אני אוהב גלידה" תומר לרצף הטוקנים [2, 1, 0] ואז לייצוג one hot:

**טרנספורמר עם דיקודר בלבד – Decoder Only Transformer:**

במודל טרנספורמר מבוסס דיקודר בלבד בלוק הדיקודר הראשון מקבל את האמבדינג והקידוד המקומי של הקלט. שאר בלוקי הדיקודר מקבלים את הפלט של בלוק הדיקודר לפניהם.

הפלט של בלוק הדיקודר האחרון מוכפל במטריצת אמבידנג משוחלפת ליצירת מטריצת לוג'יט.

על מטריצת הלוג'יט מופעלת

המודל מקבל מטריצה של ייצוג רצף הטוקנים בצורת one hot ומחזיר חיזוי בצורת לוג'יט של איזה טוקן יהיה בכל מקום ביחס בהתבסס על הטוקנים שבאים לפניו בטקסט.

במדעי הנתונים – **ווקטור הסתברות** הוא ווקטור בו מייצג את ההסתברות של מחלקה i. כל ההסתברויות בין אפס לאחד וסכום הווקטור 1.

בלמידה עמוקה – **ווקטור לוג'יט** הוא ווקטור בו מוצגות הסתברויות בטווח מינוס אינסןף עד אינסוף. פונקציית Softmax משמשת (בין היתר) להפוך ווקטור לוג'יט לווקטור הסתברות.

בעבודה שלי אשתמש במושג מטריצת הסתברות לתאר מטריצה בה היא ההסתברות של הטוקן ה ברצף להיות ה הטוקן שהמזהה שלו הוא i בהתבסס על הטוקנים הקודמים (אפס עד t). ומטריצת לוג'יט לתאר מטריצה הבנויה מווקטורי לוג'יט כמו שמטריצת הסתברות בנויה מווקטורי הסתברות.

**בלוק דיקודר Decoder Block:**

בלוקי דיקודר הם שכבות עם תתי השכבות הבאות:

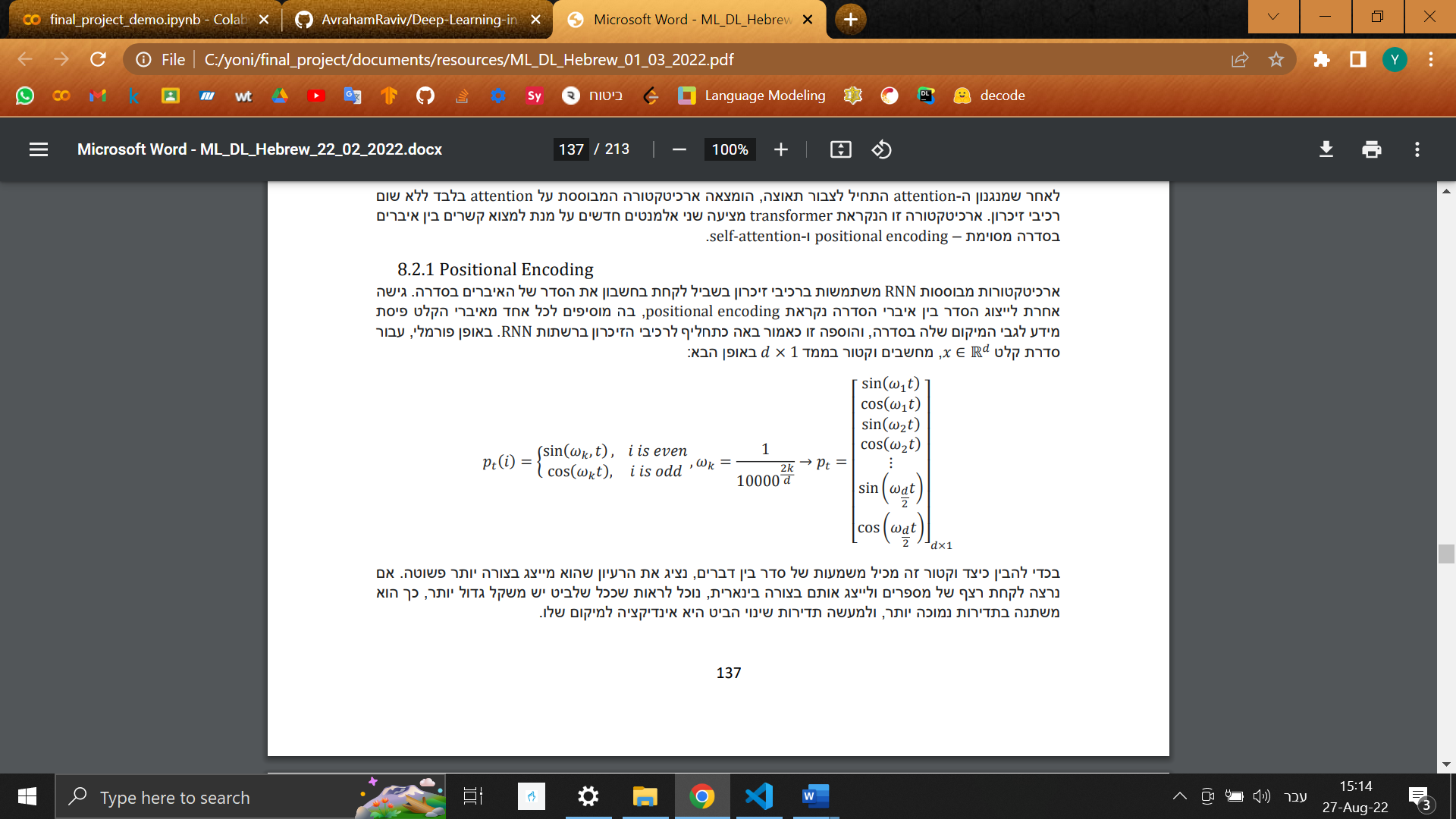
1. צומת לב עצמית רב ראשית עם מסכת הסתכלות קדימה
2. חיבור ונורמליזציה
3. רשת מחוברת לגמרי
4. חיבור ונורמליזציה

**קידוד מיקומי – Positional Encoding:**

מכיוון שהטרנספורמר אינו מתייחס באופן שונה לווקטורים במיקומים שונים בתוך רצף, יש צורך בהוספת מידע לכל טוקן בנוגע למיקומו במשפט.

שיטה בה מוסיפים לכל אחד מאיברי הקלט פיסת מידע (במקרה שלנו טוקן) לגבי המיקום שלה ברצף באופן פורמלי, עבור סדרת קלט

, מחשבים וקטור במימד x 1 באופן הבא:



**אמבדינג – Embedding:**

אמבדינג היא שכבה המקבלת רצף טוקנים ביצוג one hot וממירה כל יצוג one hot בווקטור בעל משעות סמנטית שאינו תלוי בטוקנים אחרים.

השכבה לומדת מטריצה בגודל מספר הטוקנים המוכרים למודל x שאפשר לדמיין אותה בתור טבלה בה כל שורה היא מילה שהמודל מכיר וכל עמודה היא תכונה סמנטית שיכולה להיות למילה או לצירוף אותיות בעל משמעות.

פעולת האמבדינג היא מכפלה של מטריצת הרצפים ביצוג one hot במטריצת האמבדינג.

קיימות מגוון שיטות ליצירת מטריצת אמבדינג שאינן בעזרת למידה עמוקה.

***הכפלה במטריצת אמבדינג משוחלפת:***

*את התוצאה של הדיקודר אנחנו מכפילים במטריצת האמבדינג המשוחלפת.*

*נזכור כי הכפלה במטריצה משוחלפת היא הפעולה ההפוכה להכפלה במטריצה המקורית.*

אינטואיציה:

הדיקודר מחזיר את מטריצת אמבדינג של המילים במקום השני עד המקום seq len + 1 (לפני נורמליזציה).

הכפלה במטריצת אמבדינג היא תרגום של הסתברות לאמבדינג.

הכפלה במטריצת אמבדינג משוחלפת היא תרגות של אמבדינג להסתברות.

**הפונקציה softmax:**

*softmax() מוגדרת לפי הנוסחה:*

*תפקיד הפונקציה לקחת ווקטור עם ערכים בין מינוס אינסוף לאינסוף ולהפוך אותו לווקטור שערכיו בין אפס לאחד וסכומו אחד.*

*הווקטור שהפונקציה מחזירה הוא באותו גדול של הווקטור שהפונקציה מקבלת.*

*הפונקציה מחזירה ווקטור שסכומו אחד וכל איבריו בין אפס לאחד .*

*שימו לב גם ש .*

*פעולת softmax על מטריצה בציר מסוים היא פעולת softmax על כל ווקטור בציר.*

*כלומר: פעולת softmax בציר 1 היא פעולת softmax לכל שורה במטריצה.*

***טמפרטורה - Temprature:***

*טמפרטורה היא מספר חיובי שנבחר בעת יצירת טקסט ומטרתו להשפיע על התפלגות ווקטור ההסתברות שהמודל מיצר בדגימה שאיננה לפי הסתברות מקסימלית.*

*כל איבר במטריצת הלוג'יט מחולק בטמפרטורה לפני ככה ש:*

*אינטצואיציה:*

*הקטנת הטמפרטורה גורמת לאיזון ההסתברויות של הטוקנים ככה שלכל הטוקנים תיהיה הסתברות דומה יותר והגדלת הטמפרטורה גורמת לחוסר איזון בהסתברויות ככה שלטוקנים יהיו הסתברויות שונות יותר.*

*דוגמה: מציאת ווקטור ההסתברות של ווקטור הלוג'יט [1, 1-] עם טמפרטורה של אחד, שתיים וחצי.*

*טמפרטורה איינה שימושית במהלך אימון המודל מיכוון שאם נשנה את הטמפרטורה, המודל ילמד ליצר ווקטור לוג'יט בו כל ערך מוכפל באחד חלקי הטמפרטורה.*

***מסכת הסתקלות קדימה look ahead mask***

*היא מטריצה בגודל seq\_len x seq\_len שמטרתה לגרום לכך שטוקנים לא יושפעו מהטוקנים שלפניהם.*

*כאשר seq\_len הוא אורך הרצף.*

*המסכה נוצרת בצורה הבאה:*

*def create\_look\_ahead\_mask(seq\_len):*

*answer = zeros(shape=[seq\_len, seq\_len])*

*for i in range(seq\_len):*

*for j in range(seq\_len):*

*if j > i:*

*answer[i, j] = 1*

*או לפי ההגדרה לכל איבר:*

*מסכבת הסתכלות קדימה תמיד תיהיה מטריצה ריבועית בה האלכסון וכל האיברים מתחתיו אפס וכל האיברים מתחת לאלכסון 1.*

*דוגמה: מסכת הסתכלות קדימה לרצף באורך 3:*

*אפשר לחשוב על המסכה בטור טבלה בה האיבר בשורה j ועמודה i עונה על השאלה:*

*0 אם הטוקן במקום מושפע מהטוקן במקום ו - 1 אחרת.*

**צומת לב עצמית ממוסכת בעזרת מכפלה סקלרית**

**Scaled Dot-Product** **Masked Self Attention**

המטרה של צומת הלב היא לקחת ייצוג של רצף של טוקנים ולתת לכל טוקן יצוג התלוי בטוקנים שמלפניו. כל טוקן בכל שלב במודל מיוצג על ידי ווקטור באורך (קיצור של attention dim).

נגדיר:

seq len = אורך הקלט לטרנספורמר (מספר חיובי ושלם)

*= המימד החבוי – תכונה של השכבה.*

*הן מטריצות פרמטרים הנלמדים על ידי המודל.*

*הפונקציה מקבלת מטרציה*

*הפעולה מחזירה מטריצה באותו גודל של המטריצה שהיא מקבלת.*

*הגדרה מתמטית לפעולת צומת הלב:*

*אינטואיציה:*

*המצב החבוי הוא מטריצה שמכילה רצף של ווקטורים בה כל ווקטור במקום i מיצג את המשמעות הסמנטית של הטוקן במקום i בקונטקסט של המשפט.*

*אם נדמיין כל מצב חבוי של טוקן (ווקטור בגודל ) כנקודה במרחב, הכפלתו במטריצת פרמטרים תשנה את מערכת הצירים בו הווקטור נמצא למערכת צירים שמייצגת בצורה יותר מדויקת את הקשרים שבין הטוקנים השונים. בשתי מערכות הצירים מימדים.*

*השאילתה (query) של טוקן היא ווקטור שקרוב לווקטורים שיכולה להיות להם השפעה על משמעות הטוקן.*

*המפתח (key) הוא ההשפעה של הטוקן על טוקנים אחרים.*

*הערך (value) הוא התוכן של הטוקן.*

*הפעולה יוצרת מטרציה (קיצור ל dot procut – המונח לערך שסלארי באנגלית) בגודל seq len x seq len בה הוא תוצאת המכפלה הסקאלרית בין הווקטור של שמיצג את הטוקן במקום ה i בשאילתה (לאחר הטרנספורמציה) לווקטור שמיצג את הטוקן במקום ה j במפתח (לאחר הטרנספורמציה) שמייצג את ההשפעה של הטוקן במקום i על הטוקן במקום j.*

*החלוקה של כל איבר מטריצה ב היא נורמליזציה ואינה הכרחית. אם לא נחלק הפעולה תתבצע באופן דומה מאוד והמודל יעבוד בצורה מאוד דומה. הנורמליזציה משפרת קלות את ביצועי המודל.*

*נזכור שמכפלה סקאלרית בין שני ווקטורים מייצגת את הדמיון ביניהם – ככל ששני ווקטורים יותר דומים – המכפלה הסקלארית שלהם יותר גדולה ולהפך. מכפלה סקלארית יכולה להיות חיובית או שלילית. מכפלה סקלארית לא יכולה להיות יותר גדולה מאורך הווקטור הארוך יותר בריבוע.*

*הפעולה :*

*המסכה מוכפלת במינוס מיליארד ככה שכל ערך שהיה אחד במסכה המקורית הוא מינוס מיליארד במסכה המוכפלת וכל ערך שהיה אפשר נשאר אפס ובאופן פורמלי:*

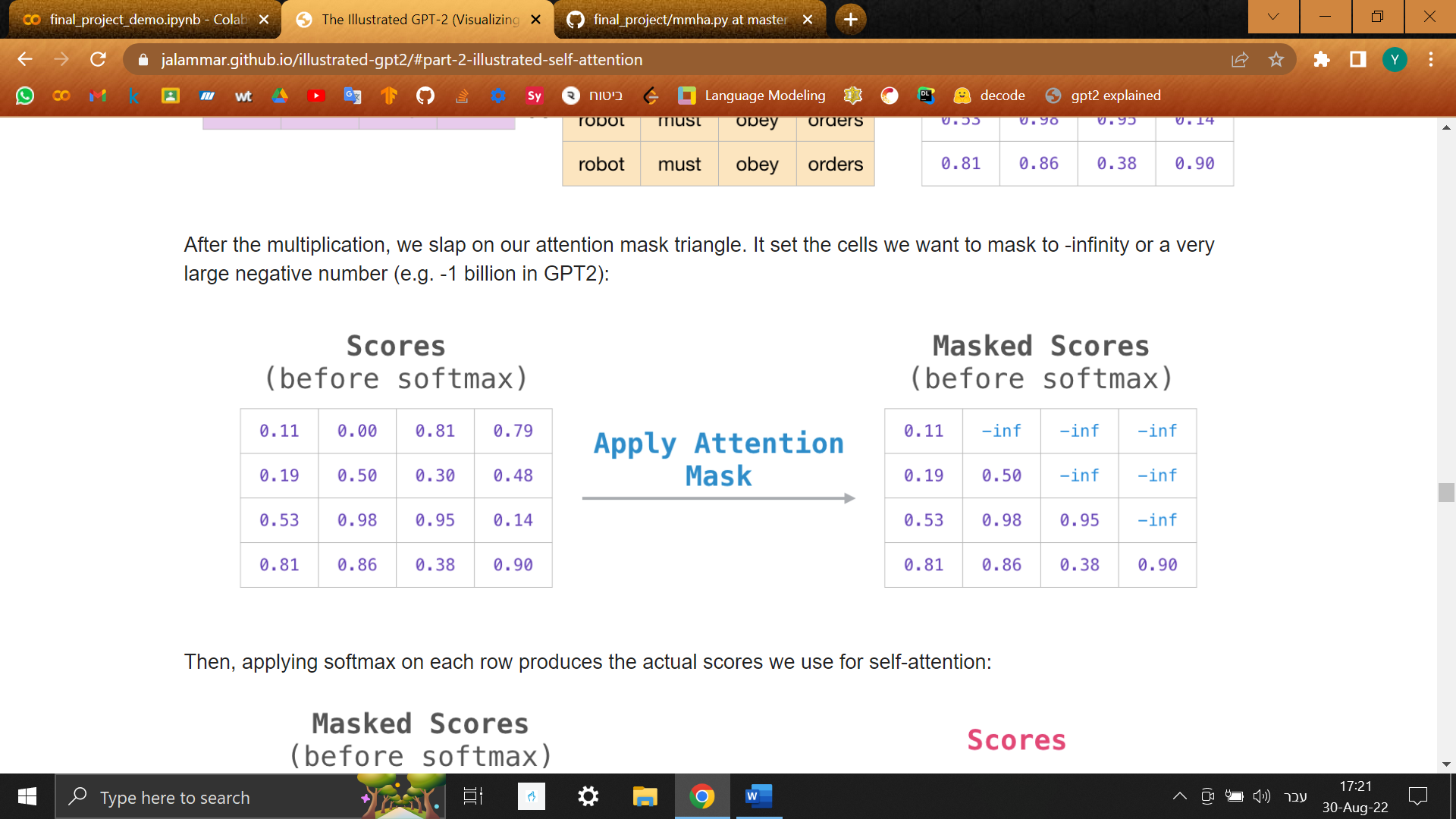
*את תוצאת המכפלה מחסרים ממטריצת הדמיון ככה שהדמיון בין*

*אינטואיציה:*

*במטריצה יש קשרים דו צדדיים בין כל הטוקנים כלומר כל טוקן משפיע על כל הטוקנים הסובבים אותו אבל זאת בעיה כי אנחנו רוצים לחזות כל טוקן מהתבסס על הטוקנים שקדמו לו בלבד.*

*אנחנו רוצים שההשפעה של טוקנים על טוקנים שבאים לפניהם ברצף תיהיה קטנה ככל הניתן.*

*המסכה גורמת להשפעה של טוקנים על הטוקנים שבאים לפניהם ברצף להיות קטנה מאוד – בערך מינוס מיליארד.*



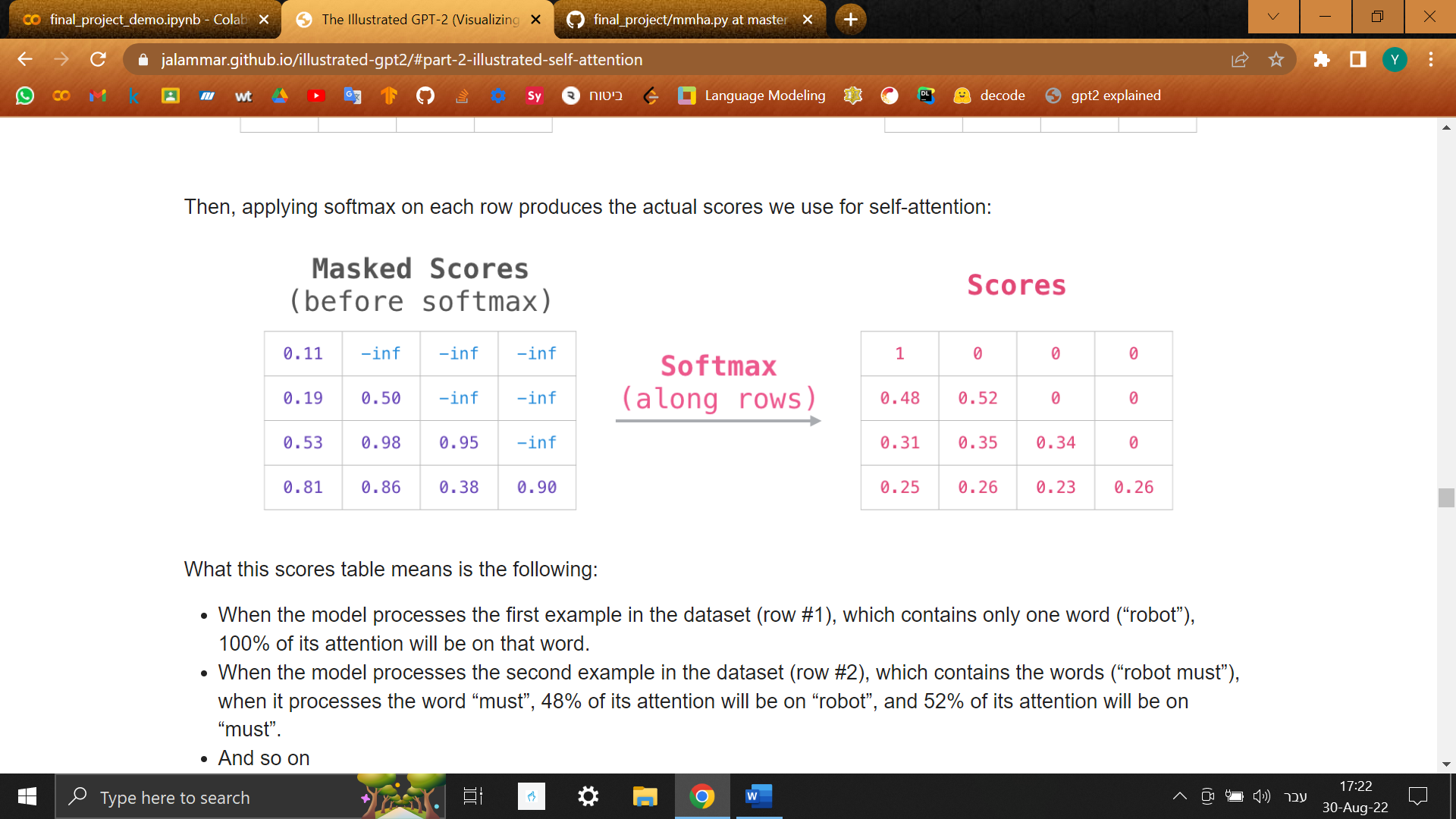
*הפעולה :*

*המטריצה ASM היא טבלה בה הוא האיבר בעמודה i ובשורה j המיצג את ההשפעה של הטוקן במקום ה i על הטוקן במקום ה j ופעולת softmax מנרמלת את הטבלה ככה שסכום כל שורה הוא 1.*

*אנחנו רוצים שההשפעה של כל טוקן על הטוקנים לפניו תיהיה קרובה מאוד ל 0.*

* *הערה: בדרך הכלל המספר הקרוב לאפס מתעגל לאפס.*

*וזאת מכיוון ש .*



*נגדיר את הערך הווקטורי של טוקן כערך שלו כפי שמיוצג במטריצה V.*

*הפעולה יוצרת מטריצת יצוג חבוי ככה ש כל טוקן מיוצג על ידי הממוצע המשוכלל של הערך הווקטורי של כל הטוקנים המשפיעים עליו.*

*אם נשתמש בדוגמה למעלה: הווקטור של הטוקן השני במטריצה יהיה 0.48 כפול הווקטור של הטוקן הראשון במטריצה ועוד 0.52 כפול הווקטור של הטוקן השני במטריצה .*

***צומת לב רב-ראשית – Multi Head Attention:***

*בהינתן אותם פרמטרים שמקבלת פעולת צומת לב ועוד פרמטר – מספר הראשים (num\_heads)*

*(המימד החבוי של המודל חייב להתחלק ב מספר הראשים).*

*הערך החבוי האחרון והערך לפני טרנספורמציה מפוצלים ככה ש המימדים שלהם משתנים מ :*

*Seq\_len x ל num\_heads x seq\_len x ()*

*ככה שלכל ראש יש ערך חבוי אחרון וערך לפני טרנספורמציה אחרים:*

*LHV\_of\_head\_i = LHV[i, :, :]*

*PTV\_of\_head\_i = PTV[i, :, :]*

*כל ראש מחשב את צומת הלב עם הערך החבוי האחרון והערך לפני טרנספורמציה שלו ולומד פרמטרים אחרים.*

*לפיצול לראשים שתי מטרות:*

1. *להקטין את זמן החישוב:*

*החישוב של צומת לב בין מטריצות קטנות לוקח פחות זמן והחישוב של כל הראשים מתבצע במקביל.*

1. *ללמוד דברים אחרים:*

*בפיצול המטריצות כל ווקטור המתאר טוקן מפוצל ככה ש:*

*vector\_size = d\_k / num\_heads*

*start\_index = i \* vector\_size*

*end\_index = (i + 1) \** *vector\_size*

*vector\_for\_head\_i = original\_vector[start\_index:end\_index]*

*נזכור שכל איבר בווקטור מייצג תכונה סמנטית של הטוקן ולכן כל ראש מתייחס לתכונות סמנטיות אחרות של הטוקנים.*

***מסכת ריפוד - Pad/Padding Mask:***

*מסכת ריפוד מאפשרת את ביצוע פעולת צומת הלב על קבוצה (batch) של רצפים באורכים שונים במקביל. לכל רצף נוסיף טוקן מיוחד - ריפוד מספר פעמים בסוף הטקסט ככה שאורך כל רצף יהיה כאורך הטוקן הארוך ברצף.*

*נגדיר = האורך של רצף הארוך ביותר, = מספר הרצפים בקבוצה (באטצ').*

*ניצור טנזור בינארי לפי הכלל:*

*. אם הטוקן במקום i+1 מושפע מהטוקן במקום j וגם הטוקן במקום j הוא אינו ריפוד ו - 1 אחרת.*

*לכל רצף במקום ה i, שכבת צומת הלב משתמשת ב במקום ב וככה נוצר מצב בו ערך כל טוקן תלוי בטוקנים שלפניו* ***וגם*** *אינם טוקן הריפוד.*

***רשת מחוברת לגמרי – Fully Connected Feed Forward Network:***

*רשת מחוברת לגמרי היא שכבה שמורכבת משלוש תת שכבות:*

*שכבה דחוסה עם גודל קלט*  *וגודל פלט feed forward depth (הייפר פרמטר של המודל)*

*אקטיבציית ReLU.*

*שכבה דחוסה עם גודל קלט feed forward depth וגודל פלט .*

***שכבה דחוסה:***

*לשכבה דחוסה שתי תכונות: גודל הקלט (n) וגודל הפלט (m).*

*היא לומדת מטריצת פרמטרים ו-ווקטור פרמטרים b בגודל m.*

*שכבה דחוסה מקבלת ווקטור שאורכו גודל הפלט ומבצעת עליה את הפעולה הלינארית:*

*הגדרה לכל איבר:*

*אינטואיציה:*

*נחשוב על הווקטור x בתור נקודה בתוך מערכת צירים, הכפלה של ווקטור במטריצה היא ייצוג של הווקטור במערכת צירים אחרת והוספה של הווקטור b היא הזזה של הווקטור בגודל וכיוון קבוע.*

***פונקציית ReLU:***

*הפונקציה פועלת על כל איבר בטנזור ונוחסתה:*

*הנגזרת של הפונקציה היא:*

***ריפוד - Padding:***

*בלמידה עמוקה אנחנו הרבה פעמי רוצים לשלוח למודל כמה דוגמאות בו זמנית.*

*עד עכשיו, אמרנו שהטרנספורמר מקבל רצף באורך שונה כל פעם אך קיימת שיטה לשלוח לטרנספורמר כמה דוגמאות בו זמנית וזה על ידי ריפוד.*

*נוסיף לסוף כל רצף טוקן מיוחד הנקרא ריפוד ככה שכל הרצפים באותו אורך.*

*לכל רצף ניצור מסכת ריפוד – ווקטור בינארי המוגדר לפי:*

***בלבול – Preplexity:***

*בלבול היא שיטה להעריך כמה המודל בטוח בתחזיות שלו כשהוא יוצר רצף.*

*הגדרה פורמלית:*

*כאשר הוא אורך הרצף שנחזה, היא ההסתברות של הטוקן במקום i (בהינתן הטוקנים הקודמים אליו).*

***דגימה – Sampling/Decoding:***

*נניח ויש לנו מודל שמקבל רצף של טוקנים ומחזירה את ווקטור ההסתברות של הטוקן הבא, יש לנו רצף של טוקנים שאנחנו רוצים שהמודל ישלים ויש לנו אורך מסוים בו אנחנו רוצים שהטקסט המיוצר יהיה.*

*דגימה היא תהליך הוספת הטוקנים לרצף. ישנן שיטות שונות לדגימה:*

1. ***דגימה לפי הסתברות מקסימלית – argmax/greedy sampling:***

*נתחילה נקבל את ווקטור ההסתברות של הטוקן הראשון על ידי הזנת הרצף המקורי למודל.*

*ניקח את הטוקן שההסתברות שלו הכי גבוהה ונוסיף אותו לרצף.*

*נחזור על התהליך עד שאורך הרצף הוא האורך הרצוי.*

*יתרונות:*

* *השיטה פשוטה וקלה לישום.*
* *אין את הסיכון של לדגום טוקן עם הסתברות ממש נמוכה.*

*החסרונות של השיטה:*

* *המטרה של רוב הטקסטים אינה להיות כמה שיותר צפוים אלא להעביר מידע.*

*אם כל מילה תיהיה המילה הצפויה ביותר בהינתן המילים שלפניה אז כמות המידע שעובר תיהיה קטנה מאוד.*

* *שיטה זו נוטה ליצר את הטוקנים שמופיעים הרבה בסט האימון של המודל (בדרך כלל מילות קישור).*
* *שיטה זו נוטה לייצר טוקנים שחוזרים על עצמם.*
* *אין שליטה: אין פרמטר שאנחנו יכולים לשנות אחרי אימון המודל על מנת לשנות את הטקסט שהמודל מייצר.*

1. ***דגימה רנדומלית מתוך הסתברות:***

*ההבדל בין דגימה מתוך הסתברות לדגימה לפי הסתברות מקסימלית היא שבדגימה רנדומלית במקום לדגום את הטוקן שהסתברותו הכי גבוהה, נדגום ברנדומליות לפי ההסתברות שהמודל חזה.*

*היתרונות של השיטה:*

* *דגימה מגוונת.*

*החסרונות של השיטה:*

* *דגימה של מילים עם הסתברות נמוכה מאוד.*
* *חוסר התאמה למטרה: המודל מאומן לחזות הסתברות כמה שיותר גבוהה לטוקן הנכון ובפונקציית המטרה אין התייחסות להסתברות של הטוקנים האחרים.*
* *תלות ברכיב רנדומלי.*

1. ***דגימה מתוך k הטוקנים שהסתברותם הכי גבוה – Top k Sampling:***

*נבחר מספר שלם וחיובי k קטן או שווה לכמות הטוקנים שהמודל מכיר.*

*בהינתן ווקטור הסתברות:*

*נמצא את k הטוקנים שהתברותם הגדולה ביותר.*

*נקבע את ההסתברות של שאר הטוקנים לאפס.*

*נחלק את ההסתברות של כל טוקן בסכום של הווקטור החדש (על מנת לקבל ווקטור שסכומו אחד).*

*ונדגום מהווקטור שנוצר.*

*יתרונות של השיטה:*

*ברוב המקרים, כל ההסתברויות השונות מ – 0.*

*יתרונות של השיטה:*

* *נותנת חסם מלמטה להסתברות של הטוקנים האופן שתלוי בווקטור ההסתברות.*
* *מאפשרת דגימה באופן שמתייחס לטוקנים שהסתברותם גבוהה בלבד.*

*חסרונות של השיטה:*

* *לא מונעת לגמרי את האפשרות לדגום טוקנים שהסתברותם נמוכה.*
* *מגדירה טוקן עם הסתברות גבוהה כ: טוקן שיש פחות מ k טוקנים שהסתברותם גבוהה מהסתברותו – הגדרה שיכולה ליצור בעיות במקרי קצה.*

*הערה: דגימה מתוך k הטוקנים שהסתברותם הכי גבוה כאשר k שווה אחד היא דגימה לפי הסתברות מקסימלית.*

1. ***דגימה מתוך הטוקנים שסכום הסתברותם <= p – Top p Sampling:***

*נבחר מספר p בין אפס ואחד.*

*אם p קטן או שווה להסתברות של הטוקן שהסתברותו הגדולה ביותר נבחר בו (בטוקן שהסתברותו הגודלה ביותר).*

*אחרת*

*ניצור ווקטור הסתברות חדש בו כל ההסתברויות 0.*

*נעבור על ווקטור ההסתברויות המקורי לפי סדר:*

*נוסיף לווקטור החדש טוקנים כל עוד סכום ההסתברות לא עולה על p.*

*ולאחר מכן נסכום את הווקטור החדש.*

*נעבור על הווקטור ונחלק את ההסתברות של כל טוקן בסכום של הווקטור החדש (על מנת לקבל ווקטור שסכומו 1).*

*ונדגום מהווקטור שנוצר באופן רנדומלי.*

*יתרונות (לעומת דגימה מתוך k הטוקנים שהסתברותם הכי גבוה):*

* *ההגדרה של הסתברות גבוהה יותר עמידה בפני מקרי קצה.*
* *השליטה בבחירת הטוקנים יותר טובה.*

*חסרון:*

* *אין חסם המנוע בחירת טוקנים בעלי הסתברות נמוכה.*

1. ***חיפוש עץ – Beam Search***

*נגדיר את ההסתברות של רצף T בו l טוקנים :*

*נבחר מספר חיובי שלם קטן או שווה לכמות הטוקנים שהמודל מכיר ונקרא לו רוחב העץ .*

*בכל שלב בחיפוש, נבחר את (רוחב העץ) הטוקנים שהסתברותם הכי גבוהה תוך מעקב על ההסתברות של הרצף.*

*לכל טוקן שבחרנו, נחזבחר את (רוחב העץ) הטוקנים שהסתברותם הכי גבוהה תוך .*

*נחזור על התהליך עד שנגיע למספר הטוקנים הרצוי.*

*לאחר מכן נבחר את הרצף שהסתברותו הגבוהה ביותר.*

*חסרון:*

*סיבוכיות זמן ריצה גדולה כשמייצריםרצפים ארוכים.*

*כאשר n הוא מספר הטוקנים שאנחנו רוצים לחזות.*

*אנחנו בעצם יוצרים עץ בו לכל הורה w ילדים ויצירה של ילדים נעשית באמצעות קיראה למודל ולכן אנחנו קוראים למודל פעמים וזמן הריצה של מודל מסוג טרנספורמר הוא ולכן זמן הריצה של הדגימה הוא בדיוק זהו זמן ריצה גדול יותר מזמן ריצה אקספנונציאלי.*

1. ***חיפוש עץ עם אילוץ על סכום ההסתברות:***

*כמו חיפוש עץ רק שבמקום לבחור את w הטוקנים שהסתברותם הכי גדולה, בוחרים את הטוקנים שהסתברותם הכי גדולה ככה שסכום ההסתברויות לא עולה על p.*

*אם המודל חוזה הסתברות שווה לכל טוקן, יבחרו טוקנים כאשר v הוא מספר הטוקנים שהמודל מכיר ולכן זמן הריצה הוא .*

***טמפרטורה - Temprature:***

*טמפרטורה היא מספר חיובי שנבחר בעת יצירת טקסט ומטרתו להשפיע על התפלגות ווקטור ההסתברות שהמודל מיצר בדגימה שאיננה לפי הסתברות מקסימלית.*

*כל איבר במטריצת הלוג'יט מחולק בטמפרטורה לפני ככה ש:*

*אינטצואיציה:*

*הקטנת הטמפרטורה גורמת לאיזון ההסתברויות של הטוקנים ככה שלכל הטוקנים תיהיה הסתברות דומה יותר והגדלת הטמפרטורה גורמת לחוסר איזון בהסתברויות ככה שלטוקנים יהיו הסתברויות שונות יותר.*

*דוגמה: מציאת ווקטור ההסתברות של ווקטור הלוג'יט [1, 1-] עם טמפרטורה של אחד, שתיים וחצי.*

**פייתון**

בחלק זה אסקור נושאים מתקדמים בתכנות ונושאים בשפת פייתון בהם השתמשתי בעבודה.

אני לא אסקור נושאים בסיסיים כגון ירושה וייבוא.

**רשימת מבני נתונים בפייתון ומבני הנתונים המקבילים בשפות אחרות:**

רשימה (list) – מערך דינמי. מושגים מקבילים: dynamic array, array list.

טאפל (tuple) – מערך שאינו ניתן לעריכה. מושג מקביל: imutable array.

מילון (dict) – מפה\מפת גיבוב. מושגים מקבילים: map, hash map.

סט (set) – קבוצה – רשימה בה אף איבר לא חוזר על עצמו. מושג מקביל: hash set.

**רמזי סוג – Type Hints:**

מערכת הטיפוסים בפייתון היא דינמית כלומר הטיפוסים בפייתון נקבעים בזמן הריצה.

פעמים רבות נרצה לדעת מה הטיפוס המוחזר בזמן קריאת הקוד ולכן יש צורך ברמזי סוג.

רמזי סוג הם הדרך המקובלת להוסיף מידע על סוג של נתונים (במקום הערות).

חשוב לציין כי רמזי סוג אינם מחייבים וכי השמת ערך במשתנה כאשר סוג הערך שונה מהרמז לסוג המשתנה אינו גורר שגיאה.

הספרייה typing מכילה כלי עזר לרמזי סוג.

**רב צורתיות – פולימורפיזם:**

עיקרון במדעי המחשב לפיו יש לקרוא למתודות אשר מבצעות את אותו תפקיד בשם זהה.

לא משנה איזה טיפוס נעביר לפונקציה, היא עושה את ההתאמות הנדרשות.

**פעולות קסם – Magic/Dunder Methods**

פעולות קסם הן פעולות שמורות בשפת פייתון שאפשר לקרוא להן בשתי דרכים:

magic(my\_object, args)

my\_object.\_\_magic\_\_( args)

פפונקציות קסם ממשות את עקרון הרב צורתיות – מחלקות רבות ממשות של פונקצייית קסם.

**קריא - Callable**

עצם קריא הוא עצם בעל פונקציית הקסם \_\_call\_\_ . הסינטקס:

My\_callable(args, kwargs)

מקביל לסינטקס:

My\_callable.\_\_call\_\_(args, kwargs)

הערה: פונקציות ועצמים קריאים הם מונחים מקבילים.

**קשטן - Decorator**

קשטן היא פעולה שמקבלת פונקציה או מחלקה ומחיזרה פונקציה או מחלקה חדשה בהתאם.

הקשטן בדרך כלל מוסיף פונקציונליות חדשה לפונקציה או מחלקה בלי לשנות את הלוגיקה.

דוגמה: קשטן שמדפיס את הטיעונים בכל קריאה לפונקציה:

def printer(original\_function):

def new\_function(args):

print(args)

return original\_function(args)

return new\_function

הסינטקסט לשימוש בקשטן:

@my\_decorator

def original\_function():

# original function's code

מקביל לסינטקס:

original\_function = my\_decorator(original\_function)

**מחלקת בסיס אבסטרקטית**

מחלקות אבסטרקטיות הינן מחלקות שלא ניתן לייצר מהן אובייקטים ומטרתן היחידה היא להוות מחלקת בסיס. במקרים בהם יש צורך להגדיר מחלקות בסיס לאובייקטים ממשיים ניתן להגדירם כמחלקות אבסטרקטיות.

לדוגמה: ניצור מחלקה אבסרקטית המייצגת צורה וממנה יורשות המחלקות ריבוע, משולש עיגול...

אנחנו רוצים ליצור עצמים ממחלקות המשנה (למשל ריבוע) ורוצים למנוע יצירת צורה כללית שאיננה אחת מהצורות הספציפיות (ריבוע, עיגול, משולש...).

על מנת ליצור מחלקה אבסטרקטית, אשתמש בקשטן abc.ABC.

**פונקציה אבסרקטית**

היא פונקציית מחלקה (mathod) של מחלקת בסיס אבסטרקטית שממומשת בנפרד לכל אחת מהמחלקות היורשות.

לדוגמה – במחלקה צורה נרצה שלכל אחת מתת המחלקות תיהיה פונקציה המחשבת שטח ובעזרת נוחסה שונה לכל תת מחלקה.

על מנת ליצור פונקציה אבסטרקטית, אשתמש בקשטן abc.abstract\_method

**פעולה סטטית**

פעולה סטטית היא פעולה ששייכת למחלקה עצמה ולא לעצם.

אחד השימושים הנפוצים של פעולות סטטיות הוא כפעולות עזר לפעולות לא סטטיות.

לדוגמה – המחלקה רובוט צריכה לבצע חישוב שאיננו תלוי בתכונות הרובוט ולכן החישוב ימומש כפעולה סטטית.

על מנת ליצור פעולה סטטית, אשתמש בקשטן staticmethod.

**מחלקת נתונים – Data Class**

הקשטן dataclasses.dataclass יוצר מחלקה ליצוג נתונים בעזרת שמות וסוגי המשתנים.

הקשטן יוצר פונקציות כגון: אתחול, יצוג כמחזורת ובדיקת שוויון.

למחלקת נתונים אפשר להוסיף פונקציות ואף לדרוס את הפונקציות שנוצרות על ידי הקשטן.

**החלק המחקרי**

**האלגוריתם:**

ביצירת טקסט בקבוצות, יצירת הטקסט נעשית כמו ביצירת טקסט "רגילה" רק שבמקום שהאלגוריתם יחשב את ההסתברות של הטוקן הבא – האו מחשב את ההסתברות של כמה מהטוקנים הבאים בהתבסס על הטוקנים שהוא כבר יצר.

**יש לקרוא את החלק התיאורטי לפני שקוראים את ההסבר על מימוש האלגוריתם.**

הקוד של האלגוריתם נמצא בקבצים: text\_generator.py, sample\_generator.py, tree\_generator.py.

**המחלקה TextGenerator** היא מחלקת בסיס אבסטרקטית (abstract base class).

משתני המחלקה:

מודל, טוקנייזר, גודל הקבוצה, טוקני ריפוד, טמפרטורה.

המחלקה מייצגת אובייקט קריא (callable) שטוען מודל שפה סיבתי מ – [hugging face hub](https://huggingface.co/models?pipeline_tag=text-generation&sort=likes) ומשתמש בו ליצירת טקסט מתוך טקסט שנקלט מהמשתמש בעזרת שיטת הדגימה שפיתחתי.

הפונקציה האבסטרקטית \_\_call\_\_ מקבלת פרומפט ומספר טוקנים ומוסיפה לפרומפט את מספר הטוקנים הרצוי.

הפונקציה: create\_prob\_matמקבלת קלט למודל בצורת מחרוזת או רצף טוקנים.

ומחזירה רשימה דו מימדית (מטריצת ההסתברות) בה prob\_mat[a][b] היא ההסתברות של הטוקן שהמזהה שלו b להיות האיבר ה a בהמשך של הטקסט הנקלט.

זאת על ידי הוספה של (גודל הקבוצה פחות 1) טוקני ריפוד בסוף הרצף.

אינטואיציה:

LHM – look ahead mask – מסכת הסתכלות קדימה.

יצירת הטוקן הבא הבאה:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| next token: |  |  |  |  |  |
| favorite | [LHM] | [LHM] | [LHM] | [LHM] | My |
| food | [LHM] | [LHM] | [LHM] | favorite | My |
| is | [LHM] | [LHM] | food | favorite | My |
| ice | [LHM] | Is | food | favorite | My |
| cream | ice | Is | food | favorite | My |

יצירת שני הטוקנים הבאים:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Next token |  |  |  |  |  |
| food | [LHM] | [LHM] | [LHM] | favorite | My |
| is | [LHM] | [LHM] | [PAD] | favorite | My |
| ice | [LHM] | Is | food | favorite | My |
| cream | [PAD] | Is | food | favorite | My |

דוגמה: אם הקלט הוא המילה "בוקר" ומזהה המילה "טוב" הוא 95 וההסתברות (לפי המודל) שהמילה "טוב" תופיע מיד אחרי המילה "בוקר" הוא 80% אז prob\_mat[0][95] יהיה 0.8 ואם מזהה המילה "לכולם" הוא 117 וההסתברות (לפי המודל) שהמילה "כולם" תופיע בצורה "בוקר" (מילה לא ידועה) "לכולם" היא 60% אז prob\_mat[1][117] יהיה 0.6.

**המחלקה TreeGenerator** יורשת מהמחלקה TextGenerator ומייצרת טקסט בעזרת חיפוש עץ בקבוצות:

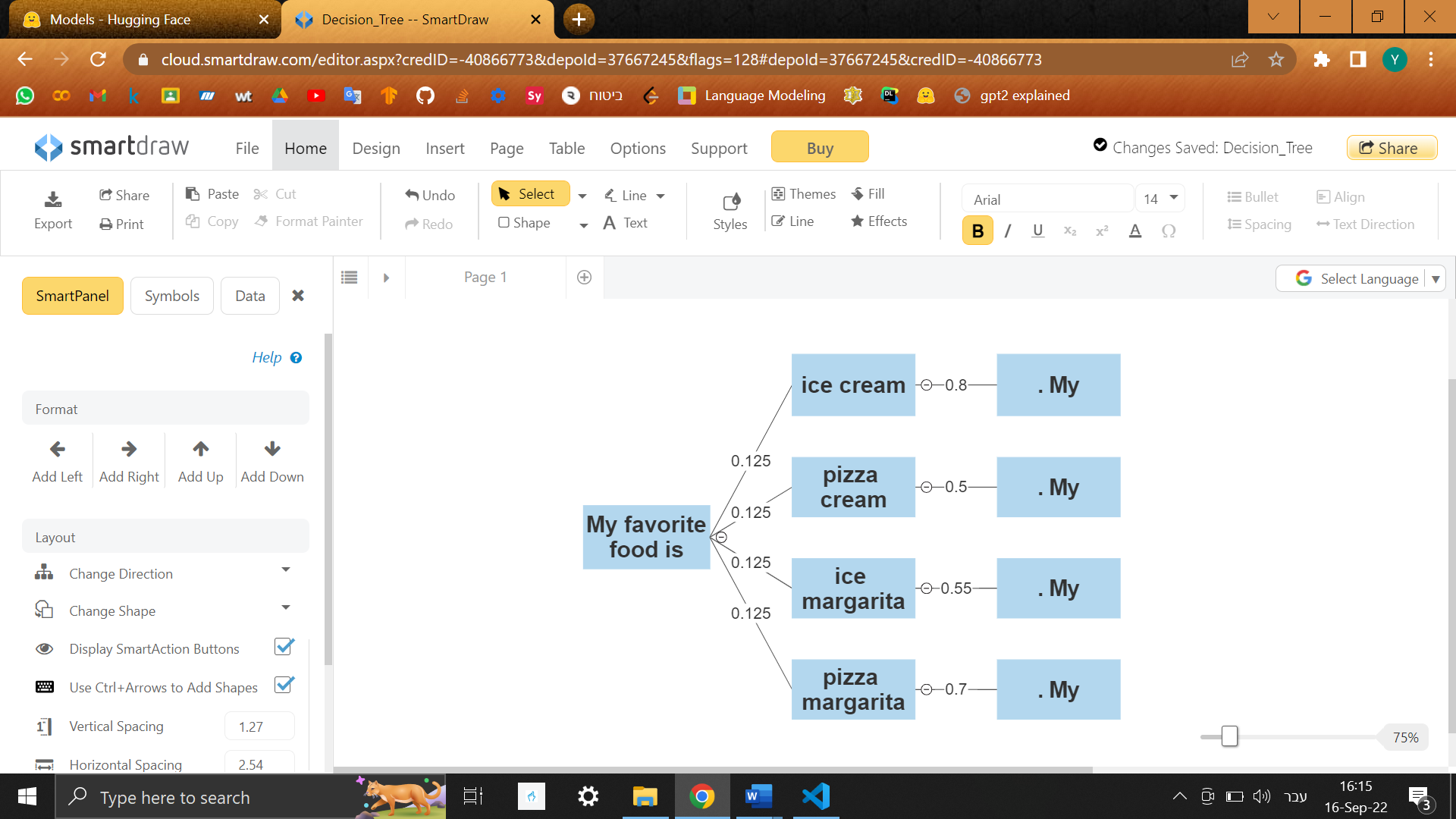
במקום לקחת את top\_k/top\_p הטוקנים שהסתברותם להיות הטוקן הבא הכי גבוהה בכל שלב - בוחרים לכל טוקן בקבוצה את top\_k/top\_p הטוקנים שהסתברותם הכי גבוהה ככה שאף טוקן לא מופיע יותר מפעם אחת – אם לטוקן מסוים ההסתברות גבוהה להיות במקום הראשון בקבוצה אז הסתברותו להיות במקומות אחרים בקבוצה נקבעת (באופן מלאכותי) לאפס.

ההסתברות של קבוצה מוגדרת במכפלת הסתברות הטוקנים בה. והסתברות של טקסט מוגדר כמפלת הסתברות הקבוצות.

דוגמה להמחשה – דגימה של 4 טוקנים כאשר גודל הקבוצה הוא 2 ו top\_pהוא 0.5 עבור הפורמפט "My favorite food is":

אם המזהה של הטוקנים pizza, ice, cream, margarita הם המספרים אפס עד שלוש בהתאם, מטריצת ההסתברות של הקבוצה הראשונה תיראה באופן הבא:

ובחירת הקבוצות תעשה באופן הבא:



הטקסט שיווצר הוא: "My favorite food is ice cream. My" מכיוון שהסתברותו הכי גבוהה.

**המחלקה SampleGenerator** יורשת מהמחלקה TextGenerator ומייצרת טקסט בעזרת דגימה בקבוצות. אסביר בעזרת דוגמה:

דגימה של 4 טוקנים כאשר גודל הקבוצה הוא 2 ו top\_pהוא 0.5 עבור הפורמפט

"My favorite food is":

אם המזהה של הטוקנים pizza, ice, cream, margarita הם המספרים אפס עד שלוש בהתאם, מטריצת ההסתברות של הקבוצה הראשונה תיראה באופן הבא:

נגריל את הטוקן הראשון לפי השורה הראשונה במטריצה ונקבל את הטוקן "pizza"

נקבע את ההסתברות של הטוקן "pizza" להיות אפס ונחלק כל איבר בווקטור ההסתברות בסכום הווקטור ונקבל שווקטור ההסתברות של הטוקן השני הוא:

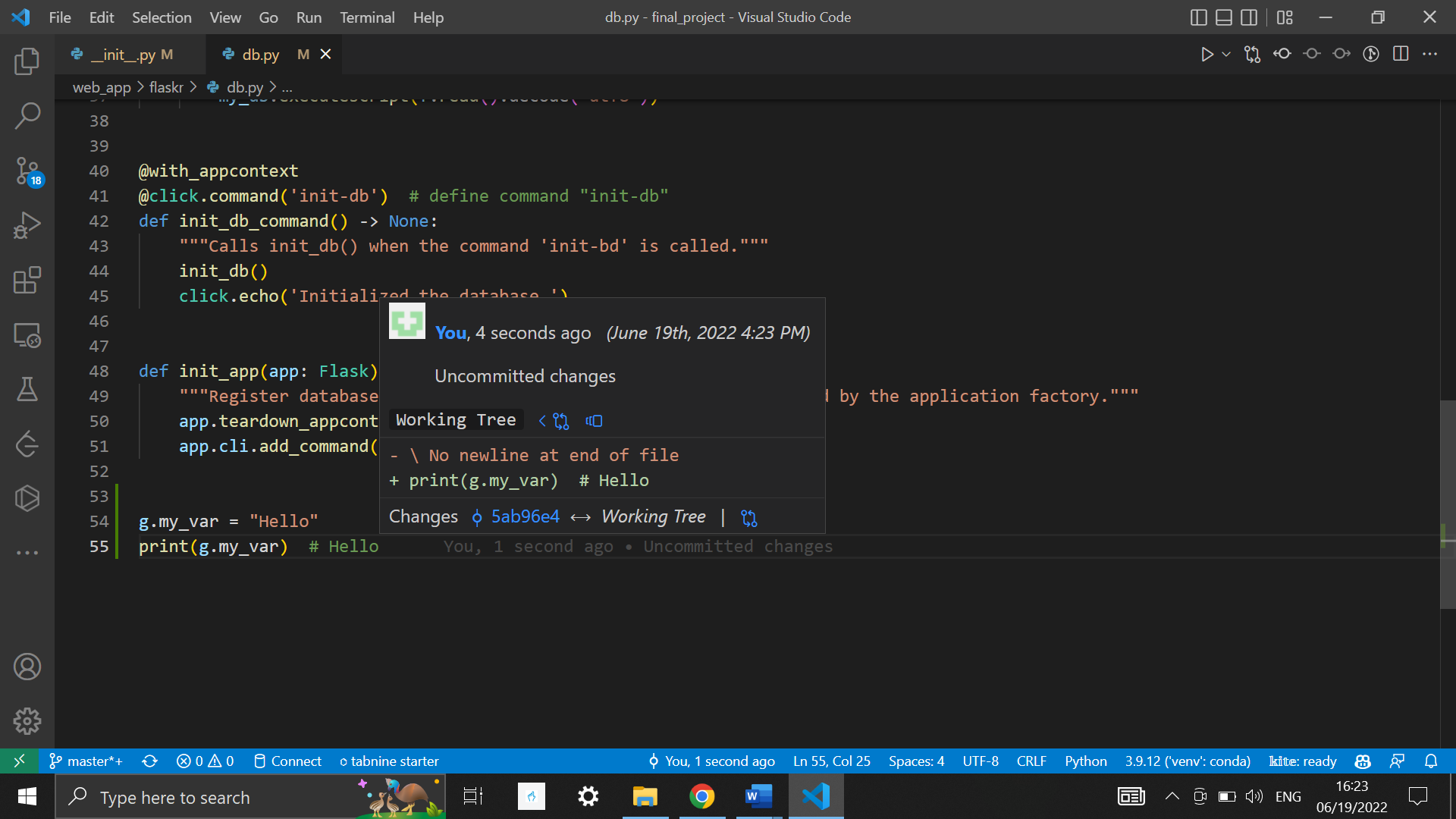
*ואז נגריל את הטוקן "margarita".*

*לאחר מכן נתחיל את אותו התהליך עם הטקסט "My favorite food is pizza margarita" וככה הלאה...*

**אפליקציית הווב**:

כל אפליקציה שמפותחת באמצעות Flask חייבת לכלול קובץ ששמו: "\_\_init\_\_.py" ובו פונקציה הנקראית "create\_app" בלבד. פונקציה זאת נקראת כשהשרת מתחיל להריץ את האפליקציה. הפונקציה מבצעת את הפעולות הבאות: יצירת עצם האפליקציה מהסוג Flask, הגדרת קונפיגורציה (למשל מיקום בסיס הנתונים), יצירת תקייה לבסיס הנתונים, יצירת בסיס הנותנים ושמירת blueprints.

המשתנה הגלובלי g מוגדר כשהמשתמש נכנס לאפליקציה. הוא ריק כל עוד לא מכניסים אליו שום דבר. אפשר לשמור בו משתנים (מכל סוג) לפי שם ולגשת בו לכל משתנה לפי שם:



בקובץ db.py נמצאות הפונקציות שאחראיות על ניהול בסיס הנתונים:

הפונקציה get\_db() בודקת אם קיים חיבור לבסיס הנתונים (עצם מהמחלקה sqlite3.Connection) במשתנה הגלובלי ואם לא, יוצרת אחד כזה ושומרת אותו ב g.my\_db ולאחר מכן (בלי קשר לתנאי הראשון) מחזירה את g.my\_db.

הפונקציה init\_db() מקבלת חיבור למסד הנתונים ומפעילה את פקודות ה SQL שבקובץ schema.sql.

הקובץ **schema.sql** מכיל את הפקודות הבאות:

אם קיימות טבלאות בשמות: completion, model, user מחק אותן.

צור את הטבלאות הבאות (כל הטבלאות בבסיס הנתונים):

user (המאחסנת משתמשים) עם העמודות:

id - מספר סידורי: שלם של המשתמש שהוא המפתח הראשי של הטבלה.

username – שם משתמש: טקסט ומיוחד.

password – סיסמה מוצפנת: טקסט (הסיסמה מוצפנת לפני שהיא נכנסת לבסיס הנתונים).

model המאחסנת מודלי למידה עמוקה עם העמודות:

Id - מספר סידורי: שלם של המודל שהוא המפתח הראשי של הטבלה.

User\_id – מספר סידורי של המשתמש שהעלה את המודל לאתר (כל משתמש יכול להעלות מספר מודלים לאתר וככה לאפשר לכל המשתמשים להשתמש בהם) בין מודל למשתמש יש קשר רבים לרבים.

model\_name – שם הקובץ של המודל: טקסט.

Created – הזמן בו נוצר המודל. ברירת המחדל היא הזמן בו המודל נכנס לבסיס הנתונים.

completions – השלמות. כל משתמש באתר יכול לבחור מודל מהמודלים הקיימים ולהכניס טקסט ולראות איך המודל משלים את הטקסט. העמודות בטבלה הן:

User\_id - מספר משתמש,

model\_id – מספר מודל,

created – הזמן בו נוצרה ההשלמה,

prompt – הקלט לאלגוריתם

answer - הפלט של האלגוריתם,

num\_tokens – מספר הטוקנים שיוצרו על ידי האלגוריתם.

Generation type – איזה אלגוריתם יצר את הטקסט? - דגימה או עץ.

top\_p, top\_k, temperature – פרמטרים של האלגוריתם.

הקובץ **auth.py** מכיל את ה blueprint auth שכולל את הפונקציה login ו register שמציגות את עמודי הכניסה והרשמה לאתר. כל אחת מהפונקציות קולטת את שם המשתמש והסיסמה מעמוד ה HTML. אם המשתמש נרשם בהצלחה הוא מועבר לעמוד ההתחברות ואם הוא התחבר בהצלחה הוא מועבר לעמוד completion.index.

הפונקציה logout מנקה את הסשין ומעבירה את המשתמש לעמוד completion.index.

הפונקציות login, logout, register מופעלות (ללא פרמטרים) כשהמשתמש נכנס לקישור auth/login, auth/logout, auth/register בהתאם.

בנוסף, הקובץ מכיל את הגדרת הקשטן (decerator) login\_required.

כשקוראים לפונקציה המקושתת בו, הוא בודק שיש משתמש במשתנה הגולבאלי (g) ואם לא, הוא מעביר את המשתמש לעמוד ההרשמה עם הודעה לפיה עליו להירשם לפני שהוא משתמש באתר. הוא מקשט את הפונקציות של העמודים הדורשים התחברות למערכת.

הקובץ מכיל גם את הפונקציה load\_logged\_in\_user שנקראת באופן אוטומטי כשמשתמש מגיע לעמודים register, login או logout הבודקת אם המשתמש שמור בסשין ואם כן שומרת או במשתנה הגלובאלי.

הקובץ model.py מכיל את ה blueprint model

הפונקציה view\_all טוענת את העמוד בו המשתמש רואה את כל המודלים שהועלו לאתר על כה.

הפונקציה get\_model\_id מחזירה את המזהה של מודל מתוך תטבלת המודלים בבסיס הנתונים בהינתן השם שלו. אם המודל לא נמצא בבסיס הנתונים – הפונקציה מוסיפה אותו.

הקובץ complition.py מכיל את ה blueprint complition.

הפונקציה index טוענת את העמוד הראשי בו מוצגות כל ההשלמות הקיימות בבסיס הנתונים.

הפונקציה create טוענת את העמוד בו המתמש מעלה כותב טקסט ובוחר מודל ופרמטרים ליצירת הטקסט. לאחר שהמשתמש לוחץ על הכפתור "Complete" המודל שהוא בחר משלים את הטקסט. ההשלמה נכנסת לבסיס הנתונים והמשתמש מועבר לעמוד הראשי.

השתמשתי ב Pytest בשביל לבדוק את כל האפליקציה. הפקודה pytest (ב command prompt מתוך התקייה final\_project/web\_app כאשר הסביבה הווירטואלית עובדת) מריצה את הקובץ conftest.py שמגדיר ומפעיל את כל הבדיקות.

**אימון מודלים:**

בתחילת המחקר, ניסיתי לאמן בעצמי מודל שפה סיבתי.

ניסיון זה כשל בעקבות מגבלה על כוח המחשוב – היה ברשותי מעבד אחד בלבד וזכרון של 12GB ולכן יכולתי לאמן רק מודלים קטנים מאוד (עד עשרות מליוני פרמטרים) שאימונם לא הניב תוצאות משמעותיות.

פונקציית המטרה שבחרתי הייתה categorical cross entropy וראיי כי ערכי פונקציית המטרה אינם מהווים שיפור לעומת חזייה של הסתברות שווה לכל טוקן ולכן החלטתי שהתוצאות שהגעתי אליהן אינן משמעותיות. בנוסף לכך, לא חל שיפור בערכי פונקציית המטרה במהלך האימון – זאת למרות נסיונות רבים לשינוי הייפר-פרמטרים.

**א – הנתונים:**

המחברת create\_models.ipynb היא מחברת ג'ופיטר שכאשר מריצים את תאיה לפי הסדר היא יוצרת מודל, מאמנת אותו ושומרת אותו בגוגל דרייב. מומלץ להריץ את המחברת דרך google collaboratory (colab) עם האצת GPU.

המחברת לפי שלבים:

1) המחברת מורידה ומיבאת את כל הספריות ההכרחיות ומדפיסה מידע על המעבד ועל גרסאות פייתון והספריות Tensorflow, Tensorflow-Text.

2) המחברת משנה את הגרעין (seed) ל-0 על מנת לשמור על תוצאות קבועות ועל יכולת שחזור תוצאות. המחברץ מגדירה משתנה device מסוג מחרוזת שיכול להיות שווה "colab" כשמריצים את המחברת על colab ו "local" אחרת. משתנה זה ישמש אחר כך למציאת מיקומי קבצים. לא השתמשתי במשתנה בוליאני כדי לאפשר יצירת אופציה נוספת במידת הצורך.

3) קביעת היפר פרמטרים: היפר פרמטרים הם מספרים שנקבעים בתחילת כל אימון מודל והערכים שלהם משפיעים מאוד על תהליך האימון. הם מתחלקים לקבוצות:

תכונות של הליך הלמידה - מספרים שמשמשים לתהליך הלמידה עצמו אך אינם תכונות של המודל שנוצר. התכונות של תהליך הלמידה אצלי הן:learning\_rate, set\_size, batch\_size

תכונות של המודל הנוצר - התכונות של המודלים שלי הן: num\_blocks,

d\_model, dff, num\_heads, dropout\_rate

תכונה משותפת – תכונה של המודל שחייבים להשתמש בה גם בתהליך האימון אצלי זאת max\_seq\_len

שימו לב שיש היפר פרמטרים שאני משאיר קבוע והם מופיעים במקום בו אני משתמש בהם ולא מגדיר בתחילת המודל-זאת מכיוון שמקובל להשאיר אותם בגודל מסוים קבוע (מכיוון שזה הערך שבסבירות גבוהה מאפשר אימון מוצלח). לדוגמה הפרמטר beta1 של ה-optimizer Adam שהערך המקובל שלו הוא 0.9.

VOCAB\_SIZE היא תכונה משותפת של הטוקנייזר והמודל אך היא שנארת קבועה כי תהליך יצירת הטוקנייזר אינו משתנה בין מודל למודל ולכן זאת תכונה שאיננה הייפר פרמטר.

שימו לב גם שהמודל לומד לזהות טקסט שעבר טוקניזציה על ידי טוקנייזר מסויים. זאת מכיוון שכל טוקנייזר נותן ערך מספרי אחר לכל מילה.

4) טעינת סט הנתונים: [סט הנתונים בו השתמשתי](https://www.kaggle.com/datasets/urbanbricks/wikipedia-promotional-articles) הוא סט של מאמרים מוויקיפדיה באנגלית שתויגו על ידי הקהילה [כמאמרים טובים](https://en.wikipedia.org/wiki/Wikipedia:Good_articles). בחרתי בו בגלל: 1. הגודל שלו (כחצי מגהבייט) שמתאים לכמות כוח המחשוב שברשותי. 2. אורך הטקסטים-הסט מכיל טקסטים ארוכים שבהם שיערתי שלמודלים שלי יהיה יתרון. 3. בקרת איכות-טקסטים שסומנו כטובים מכילים פחות תופעות כגון דברי שנאה, הסטה ודברי שקר שלא הייתי רוצה שהמודלים שלי ילמדו. 4. שפה אחידה-תמיכה בשפה אחת מפשטת את כל הפרויקט.

סט הנתונים נטען ל DataFrame בעזרת ספריית pandas והעמודה טקסט מומרת לרשימה של מחרוזות הנקראית data\_list. מודפסים נתונים בסיסיים על הרשימה.

5) יצירת אוצר המילים: בהינתן סט נתונים של טקסטים ומספר מקסימלי של מילים באוצר מילים (במקרה שלי 8192) אוצר המילים הוא רשימה של תתי המילים הכי נפוצות בסט הנתונים. תתי מילים יכולות להיות מילים או תתי מילים נפוצות (לדוגמה הסיומת ים). אל רשימה זו נכנסות "מילים שמורות": [UNK] – מילה לא מוכרת (קיצור של unknown), [START] תחילת טקסט (שונה מ-start שמציינת את המילה עצמה), [END] – סוף טקסט, [PAD] – ממלא מקום בטקסט שקצר מאורך הטקסט המקסימלי, [MASK] – מילה מוסתרת – הסבר מפורט יותר בהמשך. אוצר המילים לא מבדיל בין אותיות גדולות וקטנות. אוצר המילים נשמר בקובץ vocab.txt. אם קובץ זה כבר קיים, האוצר מילים לא נבנה מחדש אלא נקרא מתוך הקובץ. לאחר מכן כל מילה באוצר המילים ממורת לטנזור מסוג מחרוזת.

6) יצירת הטוקנייזר: look\_up\_table הוא עצם מסוג tf.lookup.StaticVocabularyTable בעל פונקציה lookup שמקבלת טנזור של מחרוזות ומחזירה טנזור של שלמים עם אותה צורה ככה שלכל איבר בטנזור:

אם האיבר באוצר מילים: "מחליפים" אותו באינדקס של האיבר באוצר המילים.

אחרת: "מחליפים" אותו באינדקס של המילה השמורה [UNK]. ("מחליפים" במירכאות כי טנזור הוא בלתי ניתן לשינוי ובעל סוג קבוע).

Tokenizer הוא עצם מסוג tf\_text.BertTokenizer בעל שתי פונקציות שאשתמש בהן:

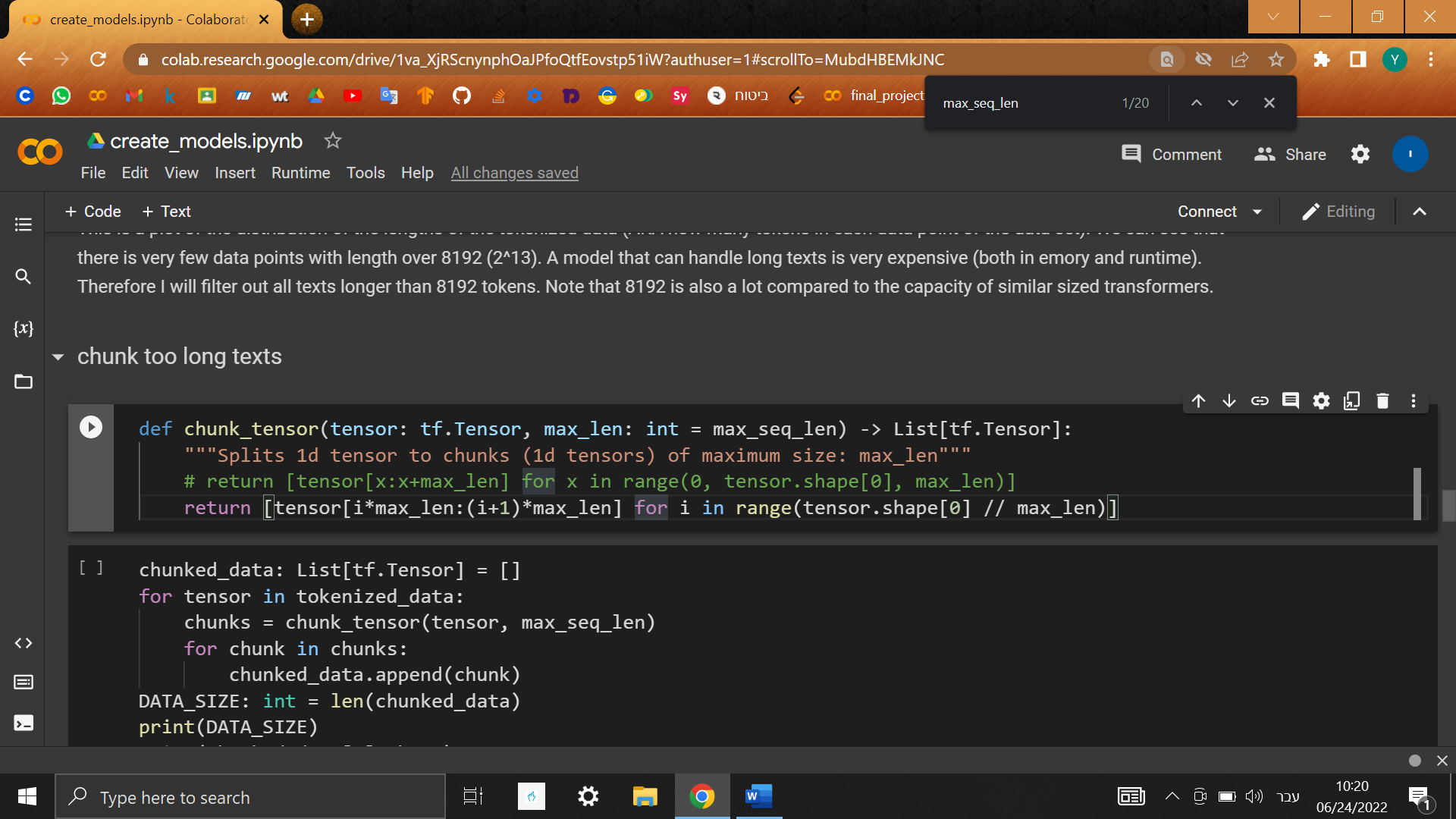
tokenize מקבלת מחרוזת, מחלקת אותה לטנזור (מדרגה 1 וסוג מחרוזת) של תתי מילים ומחזירה את הפלט של look\_up\_table.lookup() כשהיא מקבלת את הטנזור.

7) tokenization: הפונקציה tokenize\_string מקבלת מחרוזת, קוראת לפונקציה tokenizer.Tokenize עליה, ממירה אותה מ- tf.RaggedTensor ל-tf.Tensor, משנה את טיפוס האיברים מ – int64 ל-int32 ומוסיפה טוקן התחלה וטוקן סוף.

Tokenized\_data היא רשימה של טנזורים שנוצרת מקריאה ל tokenize\_string על כל איבר ברשימה data\_list. בקטע מודפס מידע על אורכי הטנזורים ברשימה.

8) חתיכת טקסטים ארוכים מדי: קבעתי את האורך המקסימלי לכל טקסט בחלק 3 של המחברת. הפונקציה chunk\_tensor מקבלת טנזור מדרגה 1 וסוג שלם ואורך מקסימלי מסוג שלם, היא חותכת את הטנזור לפי הכלל:

לכל i מ 0 עד העיגול למטה של אורך הטנזור חלקי האורך המקסימלי, החלק ה i מכיל את האיברים מ i כפוך האורך המקסימלי עד (i+1) כפול האורך המינימלי. הפונקציה מחזירה את החלקים כרשימה של טנזורים.



כל טנזור ברשימה Tokenized\_data נחתך על ידי הפונקציה והחלקים של כל הטנזורים נשמרים ברשימה של הטנזורים chunked\_data.

9) ריפוד: חלק מהטנזורים קצרים יותר מהאורך המקסימלי והמודל מקבל טנזורים שהאורך שלהם שווה לאורך המקסימלי בלבד. לכן, לכל טנזור ברשימה chunked\_data אצרף את התת מילה [PAD] שמיוצגת על ידי המספר 0.

10) חלוקה לסט אימון, מבחן ווידוי (training, test and validation): סט האימון הוא 80% מסט הנתונים, סט המבחן והוידוי שניהם 10% מסט הנתונים. כל סט מחולק ל batches ונשמר ברשימה של טנזורים בה כל טנזור הוא batch שצורתו [גודל הבאטצ', האורך המקסימלי של הטקסט] (שני הגדלים הם הייפר פרמטרים שנקבעו בחלק 3).

11) ניקוי הזכרון: בעבודה עם מחברות ג'ופיטר, המשתנים אינם נמחקים באופן אוטומטי לאחר השימוש האחרון בהם ולכן אני מוחק את כל המשתנים שאני לא אשתמש בהם בהמשך באמצעות המילה השמורה del.

**ב – המודל:**

הפרטים המלאים על חלקי המודל בחלק התיאורטי

כל המודלים שאצור הם עצמים מהמחלקה SeTransformer שיורשת מ tf.keras.Model (שיורשת גם מ tf.keras,layers.Layer) ויש לה 2 מתודות:

אתחול "\_\_init\_\_" שמתאחלת את tf.keras.Model, יוצרת קידוד מיקומים המתאים להייפר הפרמטרים של המודל ומשמש להגדרת ה encoder וה decoder ואז יוצרת את חלקי כל אחד מחלקי המודל-כולם עצמים שיורשים מהמחלקה tf.keras.layers.Layer.

קריאה: "call" שמקבלת רשימה בה שני טנזורים: פלט (inp) הוא רצף הטוקנים שהמודל צריך להשלים שאורכו כאורך הטקסט המקסימלי ומטרה (tar) הוא רצף של טוקנים באורך השווה לאורך הפלט הרצוי. ובנוסף מקבלת משתנה training שנכון כשהמודל מתאמן ושגוי אחרת. למתודה זו קוראים באמצעות המתודה predict (קריאה ישירה ל call מביאה לשגיאה).

למחלקה SeTransformer חמש תכונות שכולן יורשות בעקיפין מהמחלקה tf.keras.layers.Layer: decoder הוא עצם מהמחלקה Decoder, encoder הוא עצם מהמחלקה Encoder. embedding, dense and softmax הן שכבות הבנויות בתוך tensorflow.

המחלקות: Decoder, Encoder, EncoderBlock, DecoderBlock, PointWiseFeedForwardNetwork, MyMultiHeadAttention, ScaledDotProductAttention בנויות דומה למחלקה SeTranformer: הן יורשות (באופן ישיר, בשונה מ SeTranformer) מ tf.keras.layers.Layer ויש להן שתי מתודות: אתחול וקריאה. מתודת האתחול יוצרת את תתי השכבות ושומרות אותן במשתני מחלקה. בחלק מהמקרים היא שומרת גם הייפר-פרמטרים שהיא מקבלת, חישובי עזר ואת קידוד המיקומים. מתודת הקריאה מקבלת בנוסף למידע שמתואר בחלק התיאורטי גם מסכות שנוצרות בכל קריאה למודל.

קידוד מיקומים: בגלל שהטרנספורמר קולט מידע במקביל ולא כרצף, הוא צריך מידע נוסף על מיקומי כל תת מילה בטקסט שהוא מקבל. הפונקציה create\_positional\_encoding מקבלת את אורך הטקסט המקסימלי ואת d\_model שהוא אחד מההייפר-פרמטרים של המודל ומחזירה טנזור מהצורה # (1, , d\_modelאורך הטקסט המקסימלי) שכל אביריו בין 1 ל 1- ומסוג עשרוני (float). הפונקציה get\_angles מבצעת חישוב עזר לפונקציה positional\_encoding ולכן היא פונקציה פנימית של create\_positional\_encoding.

positional\_encoding נקראת רק מתוך מתודת האתחול של SeTransformer. התוצאות שלה נשמרות במשתנה מחלקה במחלקות Encoder ו Decoder והן שימושיות במתודת הקריאה של המחלקות הללו.

יצירת מסכות: כדי שהמודל לא יתייחס בחישוביו למטרה ולטוקן הריפוד ([PAD]), המסכות הן טנזור באותה צורה של הטנזור אותו מכסות ככה שערך כל איבר במסכה שתואם לאיבר שיש לכסות הוא 1 ושאר הערכים 0.

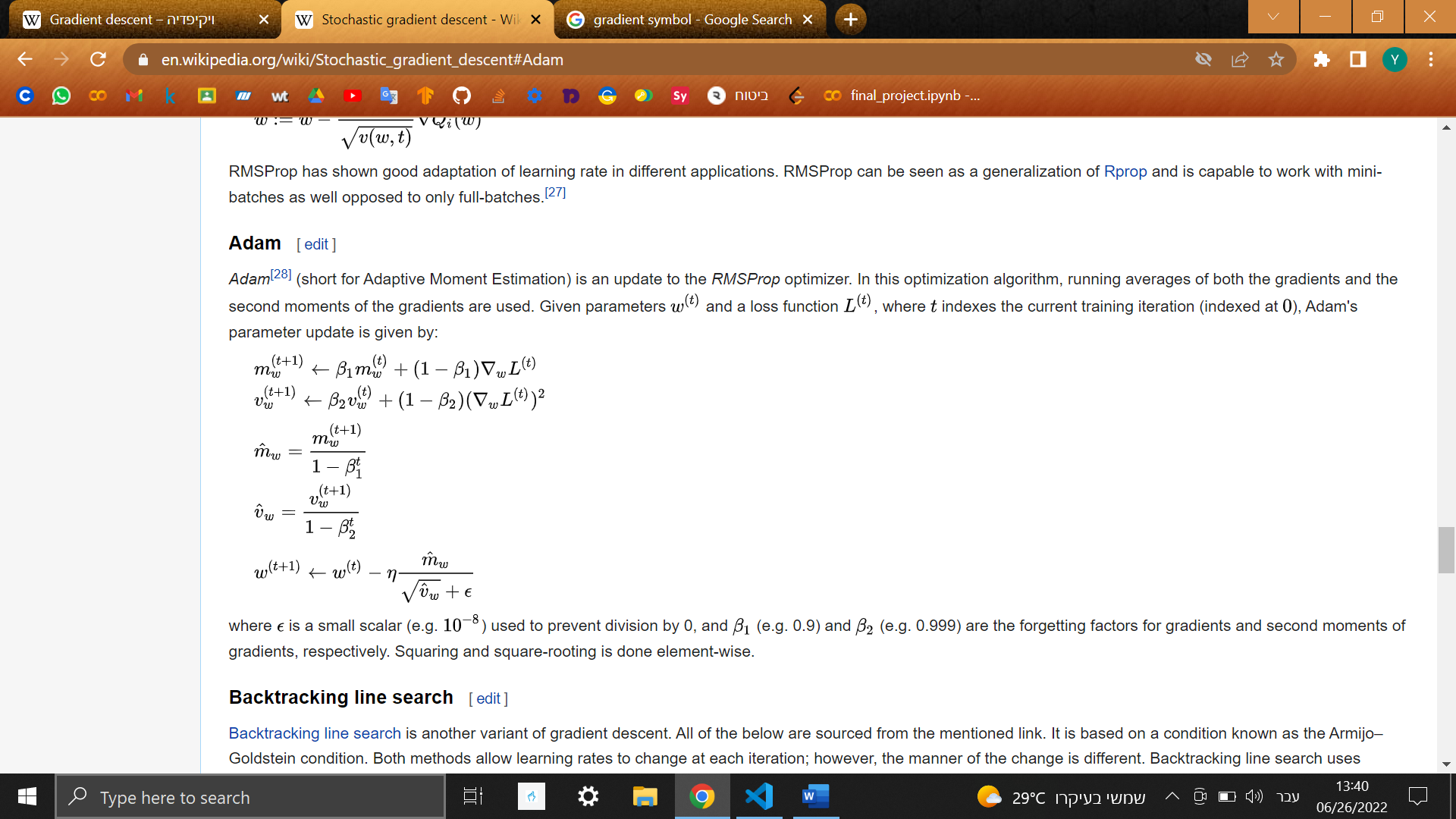
הפונקציה create\_masks מקבלת את הפלט, המטרה ואת הערך השלם שמייצג ריפוד של המודל ומחזירה מסכה אחת המכסה את המטרה ומסכה שניה המכסה את טוקן הריפוד. לפונקציה פונקציית עזר פנימית שנקראת create\_padding\_mask שמייצרת מסכה לכיסוי טוקן הריפוד שהיא גם חישוב עזר ליצירת מסכה למטרה. הפונקציה נקראית מתוך מתודת הקריאה של המחלקה SeTranformer.

**ג – האימון:**

נגדיר את המייעל להיות מייעל מסוג Adam (ראשי תיבות להערכת מומנטום הסתגלותית) עם קצב הלמידה שקבענו בחלק א 3, אפסילון ששווה לערך החיובי הקטן ביותר שמאפשר חישובים ללא סיכון של עיגול למטה ל 0 ועם בטא 1 ובטא 2 ברירת המחדל (0.9 ו- 0.999 בהתאם). למייעל מתודה apply\_gradients המקבלת 2 רשימות של טנזורים כאשר כל שני טנזורים מתאימים בעלי אותה צורה וכולם מסוג עשרוני. רשימה אחת של הפרמטרים הניתנים לאימון של המודל ורשימה של הנגזרות היחסיות של כל פרמטר ביחס לערך המתקבל מפונקציית המטרה (מספר עשרוני) ומחשבת פרמטרים חדשים למודל לפי הנוחסה (הנוסחה מחושבת לכל פרמטר בנפרד):

נקודה חשובה! כשיוצרים מודל, הפרמטרים מוגרלים באופן רנדומלי והם מספרים רנדומליים בין מינוס אחד לאחד.

כאשר: t הוא אינדקס של החזרה,  *הוא ערך פרמטר ניתן לאימון, הוא ערך המטרה (הערך שחושב בפונקציית המטרה), ו- הוא הנגזרת היחסית של הפרמטר היחס לערך המטרה. זאת הנוסחה:*



נשתמש בפונקציית המטרה: sparse categorical crossentrop.

הפונקציה מחשבת את השוני בין טנזור הסתברות בו האיבר במקום i מייצג את ההסתברות לתוצאה עם אינדקס i לבין האינדקס של התוצאה הנכונה. מטרת המודל שלנו היא לחזות הסתברות כמה שיותר גבוהה לתוצאה הרצויה. ערך גבוהה של הפונקציה שקול לחזייה לא מוצלחת ולהפך ולכן מטרת המודל לצמצם את ערך הפונקציה.

הפונקציה מוגדרת עבור הסתברות של התוצאה הרצויה p:

כשההסתברות שואפת ל 0, ערך הפונקציה שואף לפלוס אין-סוף וכשהיא 1, ערך הפונקציה 0. הפונקציה מתעלמת מההסתברות של כל תוצאה אחרת.

לאחר מכן, ניצור את המודל עם הייפר-הפרמטרים שבחרנו בחלק א 3.

הפונקציה train\_step ממשת את אלגוריתם ה Stochastic gradient descent עם מייעל הערכת מומנטום הסתגלותית.

כל מודל למידת מכונה הוא בעצם פונקציה בעלת הרבה פרמטרים (במקרה שלי עשרות מליונים) שניתן לגזור בכל נקודה. כאשר כל קלט תקין למודל הוא נקודה במרחב הקלט האפשרי. גם פונקציית המטרה היא פונקציה מתמטית גזירה ולכן אפשר למצוא את הנגזרת היחסית של כל פרמטר במודל ביחס לפלט. הנגזרת היחסית של פרמטר ביחס לפלט מייצגת את השפעת הפרמטר על הפלט (גודל וכיוון).

Tape הוא אובייקט מהמחלקה tf.GradientTape שעוקב אחרי הנגזרות של כל הפעולות בעל פונקציה gradient שמחשבת את הנגזרת היחסי בין כל שני משתנים.

הפונקציה train\_step מקבלת את הפלט והקלט הרצוי בצורת רצץ של טוקנים, קוראת למודל ומחשבת את פונקציית המטרה תוך כדי מעקב על הנגזרות, מחשבת את הנגזרת היחסית של ערך פונקציית המטרה ביחס לפרמטרים של המודל ואז משנה את הפרמטרים של המודל בעזרת המיעל. הפונקציה מחזירה את ערך פונקציית המטרה. היא מקושטת בעזרת הקשטן tf.function.

הפונקציה validate מקבלת באטצ' מסט הוולידציה או המבחן ומחשבת את ערך פונקציית המטרה הממוצע.

הפונקציה check\_points מקבלת מיקום רצוי מסוג מחרוזת, מודל מסוג SeTransformer וערך פונקציית המטרה הממוצע על סט האימון, הוולידציה והמבחן מסוג עשרוני. היא שומרת את המודל במיקום הרצוי ומוסיפה שורה לקובץ expirement.csv שכל שורה בו מייצגת מודל אחד שלו ובשורה יש מידע על היפר הפרמטרים של המודל כמו גם על ערך פונקציית המטרה שלו על סט האימון והוולידציה. אם קיים מודל שמור מאותו תאריך, הפונקציה מוחקת אותו וזאת מכיוון שהמודל הוא בהכרח גרסה מוקדמת יותר של המודל הקיים שביצועיה נמוכים.

לולאת האימון:

יוצרת רשימות ריקות של ערכי פונקציית המטרה על סט האימון והוולידציה בכל חזרה מסוג עשרוני.

לכל מספר חזרה בטווח (0 עד מספר גבוה) (עד שהמחברת תנתק מהאתר או שפונקציית המטרה תתחיל לעלות):

יוצרת רשימות ריקות של ערכי פונקציית המטרה על סט האימון והוולידציה בכל באטצ' מסוג עשרוני.

לכל באטצ' בסט האימון:

מחלקת את הבאטצ' לסטים, קוראת לפונקציה train\_step על כל סט ועוקבת אחרי ערך פונקציית המטרה. אם מספר הבאטצ' מתחלק ב 8 = כמות הבאטצ'ים בסט האימון-80% מסך הבאטצ'ים חלקי כמות הבאטצ'ים בסט הוולידציה-10% מסך הבאטצ'ים:

קוראת לפונקציה validate על הבאטצ' ועוקבת אחר ערך פונקציית המטרה.

ערך פונקציית המטרה של החזרה מוגדר לממוצע של ערכי פונקציית המטרה על כל באטצ'.

אם ערך פונקציית המטרה על סט הוולידציה לא משתפר:

האימון מופסק.

אחרת אם מספר החזרה מתחלק ב 10: שמור את המודל בעזרת הפונקציה

מוצג גרף של ערך המטרה לפי חזרות.

ערך פונקציית המטרה על סט המבחן מחושב על ידי הפונקציה validate והמודל נשמר.

**הדמו:**

הקובץ project\_demo.ipynb הוא מחברת ג'ופיטר אותה שאותה המשתמש פותח דרך colab .

במחברת המשתמש בוחר מודל מאומן מראש, ופרמטרים לדגימה (top k, top p, number of goups, group size) מכניס טקסט ורואה את הטקסט שהמודל חוזה בשיטת הדגימה שאני מציע בעבודה לעומת שיטות קודמות.

המודל המאומן מראש יכול להיות כל מודל שהועלו ל hugging face ויורש מהמחלקה [AutoModelForCausalLM](https://huggingface.co/docs/transformers/main/en/model_doc/auto#transformers.AutoModelForCausalLM). [בקישור הזה](https://huggingface.co/models?pipeline_tag=text-generation&sort=likes) אפשר למצוא רשימה מלאה של כל המודלים שמקיימים את הדרישות הללו.