עיבוד שפה טבעית (20225211) ב2024-2025 - תרגיל 4 חלק 2 (4 נק^י) - עיבוד שפה טבעית

שלום לכולםן!

כעת נתבונן קצת יותר לעומק בשיטות ליצירת אוצרות מילים (vocabularies) מתחת לרמת המילה (subword). נאסוף קורפוס, נאמן טוקנייזר לעברית שמתאים ל-BERT, ונבחן את תכונותיו.

בסיום העבודה, הגישו **קובץ pdf** הכולל את כל הפלטים מכל התאים. דרך אפשרית אחת להשיג קובץ pdf היא על-ידי הדפסת הדף (מתוך colab, לא מהדפדפן) ולבחור "מדפסת" ישירה ל-bdf.

שימו לב שכל המידע הדרוש מופיע בתדפיס ה-pdf --- לא יתקבלו ערעורים על-בסיס חיתוך. אין לחרוג ממבנה התאים ואין לענות במקומות שאינם מוגדרים לכך. אין למחוק הערות שמסמנות את מיקומי קטעי הקוד שלכםן ואין לערוך קוד שנמצא מחוץ לתחומים המוגדרים. סעיפים שחורגים מהנחיה זו לא ייבדקו.

ניתן להגיש בצוותים של עד שלוש.ה סטודנטיםות. לא יהיו הנחות לצוותים קטנים. אפשר לחלק את פתרון התרגיל בין חברי.ות הצוות כרצונכםן. אין להסתכל על קוד של צוותים אחרים ואין להקריא קוד לצוותים אחרים.

או כל סייען מתקדם אחר. Gemini אין להשתמש ברכיב ג'מיני

מועד ההגשה הוא דקה לפני תחילת השיעור ב-13/1. לא יינתנו הארכות.

בהצלחה,

יובל --

התקנות

דבר ראשון, נתקין את החבילות הנדרשות, כולן מבית האגינגפייס.

```
!pip install datasets -q
!pip install tokenizers -q
!pip install transformers -q
```

נייבא חבילות. את os נצטרך כדי לכתוב את הדאטאסט שלנו לטובת האימון, וכדי לכתוב את אוצר המילים של הטוקנייזר לאחר האימון כדי שיוכל להיטען לאובייקט הקורא אותו.

```
import os
import datasets
from tgdm.notebook import tgdm
```

החבילה datasets מכילה קורפוסים לא-מתויגים רבים, בדיוק מהסוג שאנחנו צריכים לטובת אימון טוקנייזר תת-מילי. נשתמש ב- <u>oscaR</u>, פרויקט נקיון ותחזוקה של מידע ממאגר הגריפה הגדול commoncrawl התומך ב-166 שפות, וביניהן עברית.

נבקש ממנו את מחיצת האימון (היחידה שקיימת) לעברית ונספק את הפרמטר streaming כדי שנוכל לדגום חלק קטן ממנה בלי להוריד הכל לזכרון של קולאב (המון זמן). 200,000 מסמכים יהיו די והותר.

```
dataset = datasets.load_dataset(
   'oscar',
   'unshuffled_deduplicated_he',
   split='train',
   streaming=True,
   trust_remote_code=True)

sampled_ds = []
for i, doc in tqdm(enumerate(dataset)):
   if i > 200_000:
        break
   sampled_ds.append(doc)
```

Start coding or generate with AI.

היעלה על הדעת שנוריד מידע ולא נתבונן בו? הבה נבחן את המסמכים האחרונים שהורדנו.

sampled_ds[-10:]

→ [{'id': 199991,

'text': יודער בארץ ובעולם של הסרט החדש' "Star Wars The Force Awakens" ולכבוד הסרט גוגל בנו משחק קצרצר של מלחמת הכוכבים הפועל המחדש המיוחד שיוצג לכם באתר במחשב מהפלאפון ולאחר ביצוע כיול קצרצר ופשוט תוכלו להדוף יריות ולבסוף גם להלחם בח\.בתוך דפדפן הכרום Storm שיוצג לכם באתר במחשב מהפלאפון ולאחר ביצוע כיול קצרצר ופשוט תוכלו להדוף יריות ולבסוף גם להלחם בח\.בתוך דפדפן האטהח\.תניות: גוגל, דפדפן, כרום, ח\.לצערי המשקח מאוד מאוד קצר ומכיוון שמדובר במשחק דפדפן באיכות לא רעה ושליטה שניעשת ממכשיר אחר דרך הרשת יש מדי פעם קצת האטהח\. (Chrome, Lightsaber, Star Wars, Storm Trooper, The Force Awakens'), מלחמת הכוכבים, משחק, סרט ('id': 199992,

התרגלנו לחשוב שיש לנו סוג של בעלות על האדם האהוב עלינו. לפעמים מספיק שנראה את הבן זוג שלנו מסתכל על אישה אחרת, והקנאה מתעוררת. האמירה הבאה אולי' :'text': לפואה המזרחית גישה מעניינת ואפקטיבית מאוד, לתופעות גיל המעבר. המחלקה לרפואהח\תפתיע אתכם: קנאה אינה שייכת לאהבה! אם תאמצו זאת כמוטו, חיי האהבה שלכם ישתנו לחלוטין סינת ויפנית במרפאות אלטרנטיבה מציעה טיפול ייחודי משולב לתופעות המטרידות של גיל המעבר: גלי חום, הועות לילה, חרדה, מתח ואי שקט, דפיקות לב מואצות, הפרעות שינה ודיכאון. אנשי עסקים מצלחים שהכסף מצוו בכיסם מגיעים לפעמים לצומת דרכים. הכסף מסתבר לא עונה על הכל והתשובה עשויה להמצא בדווקא במקומות החיבור לשפע חומריח\רווזניות ואנשי עסקים הצער\רהודניות החדשה של היום, טוענת שאין סתיה בין רוחניות וכסף. בהחלט אפשרי ליהנות משני העולמות. כידצ עבודה דרותנית כולת לתמוך בחיבור לשפע חומריח\רווזניות ואנשי עסקים מנדלה היא יצירה שצומחתח\יודע לשחרר את הישן ולפנות מקום לחדש: העלים היבשים נושרים ועלים ירוקים חדשים יצמהו במקומם, בנוא העת. רק האדם לעתים נצמד לישן, מפחד משינויים מדלה היא מיר מפירושה מעל מקודש. היא מלווה דתת שונות ואף שבטים ותרבויות קדומות עוד מתרו מעל פנימי, אשר גדלה ומתפתחת לעוד ועוד מעגלים סביבו, עד אין סוף. מנדלה היא מירם בתחומים אישיים. כיצד לקבל את השינויים כחלק בלתי נפרד מהחייםח\מהרות את יחסנו אליהם, הן בתחומים מקצועיים והם בתחומים אישיים. כיצד לקבל את השינויים כחלק בלתי נפרד מהחיים אליהם, הן בתחומים מקצועיים והם בתחומים אישיים. כיצד לקבל את השינויים כחלק בלתי נפרד מהחיים אליהם, הן בתחומים מקצועיים והם בתחומים אישיים. כיצד לקבל את השינוים כחלק בלתי נפרד מהחיים לום: "לוו לינו לשנות את יחסנו אליהם, הן בתחומים מקצועיים והם בתחומים אישיים. כיצד לקבל את השינוים כחלק בלתי נפרד מהחים או ליום בתחומים אישים ליום למום לום בתחומים מעצר מדיה מחור מתוחים במום למום למום בתחומים מעובר מתחים מתחומים מתחבר מתחבר מתחים מתחבר מ

'text': משפחה, חברה, משפחה, חברה, תחומי עיניין אישיים . לא פעם אנו מוצאים את עצמינו "נקרעים" בין ה"עולמות" הנ"ל ולא פעם' : לא פעם' 'נינין אישיים . לא פעם אנו מוצאים את עצמינו "נקרעים" בין ה"עולמות הסיפוק אישי, או מקצועי ועוד (שפוז במסגרתח). התוצאה היא הזנחה של חלק מתחומים, תשלום מחירים גבוהים בבריאות , בקשר עם הילדים , בקשר עם הבן זוג , ברמת הסיפוק אישי, או מקצועי ועוד במעגלי חיים השונים (במסגרתח). מושגת תוך סיפוק הצרכים האישיים, המקצועיים והמשפתיים. סיפוק צרכים אלו מתאפשר ע"י ניהול אפקטיביים לניהול המשימות ולניהול זמן , נאתר יחד את הדרך לאיזון האנרגיות ('id': 199994,

הנוסע המוסלמי אבן בטוטא (1377-1304) ביקר בטורקיהח\דף הבית >> סרץ ארכיאולוגי\והיסטורי של אתרים מרכזיים יוונים ורומיים במחוזות טורקיה וליאר אתה כאזור המופלא שבעולם, אכן מדינה מרתקת, ייחודית גן עדן ארכיאולוגי, נופים וחופים מהיפים בעולם. החל משנת 1998 לאחר לימודי הארכיאולוגיה ביקרתי, סיירתי האזור בו שוכנת כיום טורקיה המודרנית היה ערש הולדתן של תרבויות חשובות, בהן האימפריותח\וצילמתי יותר מ- 20 פעמים, היתה לי הזכות לצלם מאות אתרים ארכאולוגים ואלפי מונומנטים הביזאנטית והעות\'מאנית. בשל המיקום האסטרטגי של טורקיה, בנקודת המפגש בין אסיה ואירופה, התרבות הטורקית היא מיזוג מיוחד במינו של המסורת המערבית והמודחית. יש לה גם חשיבות ה-18 וה-13 לפנה"ס. הפריגים ירש את החתים עד שהקימריםח\. גיאו-פוליטית אסטרטגית משמעותית הראשונה באזור הייתה האימפריה החתית שהתקיימה בין המאות ה-18 וה-13 לפנה"ס. הפריגים ירש את ההחתים עד שהקימריםח\. גיאו-פוליטית אסטרטגית משמעותית החחשובות שבהן היו לידיה, קאריה וליקיה, כולן בחלקה המערבי של אסיה הקטנה, ודיונים החלו מיישבים אותה לקראת סוף האלף השני לפנה"ס. האור כול נכבש על ידי הממלכה האחמנית במאות ה-16 וה-15 לפנה"ס, ולאחר מכן על ידי אלכסבדר הגדול בשנת 334 לפנה"ס. מתקופה ההלניסית חלקו מספר ממלכות את שטחה של אסיה הקטנה ובמאה ה-19 לפנה"ס החלו הרומאים בכיבושה והשלימו אותו בשנת 65 לפנה"ס לערןטורקיה היא רפובליקה זמוקרטית, חילונית, מוסעפא כמאל אטאטורק בנה את המערכת הפוליטית של האומות המאוחדות, ושל ארגון המדינות האסלאמיות, והצטרפה לנאטו ב-1952. מאז למנתה עם החברות המיסדות של האומות המאוחדות, ושל ארגון המדינות האסלאמיות, והצטרפה לנאטו ב-1952. מאז למנחה בחברות לאיחוד האירופי לאחר נפילת האימפריה העות\'מאנית. טורקיה מתח החברות המיסדות של האומות המאוחדות, ושל ארגון המדינות האסלאמיות, והצטרפה לנאטו ב-1952. מאז למנחה עם החברות המיסדות של האומות האוחדות, ומילת האטלמית, והצחות לצאות העות לאימות, והאחודות לאחר נפילת האומות לאום להצטרם החברות המיסדות של החברות המיסדות של החברות המיסדות המיסדות של החברות המיסדות המיסדות של החברות המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיח המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיסדות המיחדית המיחדים המיסדות המיסדות המיסדות המיסדות המיסדות המיחדים ב

{'id': 199995,

'text': '.חוב או אם זה למרחב, בכל סוג של פרויקט מתקן החשמל דורש שיטות תכנון מתקדמות אשר התפתחו עם השנים מליבה יציבה ומוכחת של ידע, ניסיון ועבודת צוור.' בשנים בארץ ובעולם שימוש בטכנולוגיות חדשות לצורך קביעת סטנדרטים חדשים במתקני החשמל אותם מתכנן המשרד ושילובם בחזית ההתפתחות הטכנולוגיות העצומה שחלה בהקמת פרויקטים מורכבים בארץ ובעולם "בשנים האחרונות".'},

{'id': 199996,

ילצארי: 'מרות במחירת מקרר הוא בחירת הנפח המתאים. מקרר גדול מידי יגרום לבזבוז אנרגיה ומקרר קטן מידי יהיה מוגבל ביכולת האחסון, 600-600 ליטרים' 'אמדית חדשות בפנח מתאים למשפחה הממוצעת. ישנם שני סוגים של מדידות נפח: "מדידות אלו כוללות את כל הנפח של המקרר משמע ללא המגירות ובצבעים שונים, האפשרויות בחירה באתר שלנו גדול מאוד וכאן תוכלו למצוא כל מקררח\. - מדידות אלו כוללות את המדפים והמגירות ולכן זהו הגפח נטו מקררים יכוללוגיה זו כבר ישנה ואט אט מוחלפת ע"י הטכנולוגיה מותר - De-frost שהמקרר מיצרו עם ציפוי מוחלפת ע"י הטכנולוגיה זו מיצרת קרה בדפנות הפנימיות של המקרר. קרה זה מצטבר עקב האויר הקר והלח שהמקרר מיצר ועקב כך צריך להפסיק את פעולת המקרר אחת לחודש על מנת להפשיר את טכנולוגייה זו מתקדמת יותר וקריכת בל המקררים החדישה ביותר, לא צריך להפסיק את פעולת משום שלא מצטבר קרח בדפנות אקרח המדובר. טכנולוגיית מיצרת קרו במקרר מנת של יצטבר קרח בדפנות, אוויר קר ויבש פוגע בצורה משמעותית יותר באוכל שבמקרר מאשר לאסכנולוגיה זו יש גם חסרון: טכנולוגיית המקרר חדש, עדיף לבחון את הדירוג האנרגטי של כל מוצר נקבר על ידי משרד האנרגטי של כל מוצר נקבע על ידי משרד האנרגיה לפני שהמוצר נכנס לארץ עוד במכון התקנים. צריכת אנרגיה ושל כו מוצר מקרר אור מקרר שור מורים מקרר וותר מקרר יותר אפקטיבי בצריכת אנרגיה מובר אור מורים מלאכותית. במצעות חיישני טמפרטורה, מזהה המערכת שינויי טמפ\' ומזרימה\'.מקרר קטן, דב זה לא אומר שמקרר קטן שורר בהב- זותר אפקטיבי בצריכת שוריה לשמור על טמפ מלאכותית. באמצעות חיישני טמפרטורה, מזהה המערכת שינויי מפא' לפינוי המוצר הישן מביתם ללא תוספת תשלום עפ"י החוק, לחצו לפרטים המלאים לפינוי המוצר הישן מביתם ללא תוספת תשלום עפ"י החוק, לחצו לפרטים המלאים לפינוי ממוצר הישן מביתם ללא תוספת תשלום עפ"י החוק, לחצו לומיר ("ול" במטרה לפינוי ממוצר הישן מביתם ללא תוספת תשלום עפ"י החוק, לחצו לפרטים המלאים לפינוי ממוצר השום לפינוי ממוצר הישן מביתם המלאים לפינוי ממוצר הישן מביתם ללא תוספת תשלום עפ"י החוק, לחצו למלים ליטר ("מו"ם "מוצר" לפינוי ממוצר השום לפינוי המוצר הישן מביתם המאבר הישן מביתם המצר הישן מביתם המוצר הישור לישור לא

'text': 'תומר עליוןח\אביב / סתיו נעליים ללא שרוכים לבן / שחור 2018 6534228 בגדי ריקוד גברים נעלי נוחות סוויד' וימא בירום (עליים ללא שרוכים לבן / שחור צפספס ביותר במרחב ההנעלה. לא רק בגלל אופנה אלא גם מעשי. אתה אף פעם לא יכול לפספס נעלי זמש ('id': 199998,

כדי לסייע לכם להתאים את השיש לבחירות השונות שצריך לעשות במסגרת תהליך של עיצוב מטבחים, נועד מאמר זה. השתמשו בו בתבונה והשיגו ממנו את המרב. אתם בהחלט' :'text': דעו כי בכל הנוגע לבחירת שיש,ח\.יכולים לקחת את מה שמועיל לכם ולהתעלם ממה שלא נראה לכם רלוונטי, אך דעו כי הבחירה הנכונה היום, היא שתאפשר לכם ליהנות מהשיש מחר ובכלל אין בחירה נכונה ובחירה לא נכונה, משום שהבחירה הינה אינדיווידואלית - מה שנכון לאחד לא נכון לאחר ולהיפך. הבחירה תלויה בין היתר באהבה שלכם לבישול ואפיה וכפועל יוצא תדירות נשאלת השאלה החשובה מתי נכון לבחור את השיש והאם לתקופה בה מתבצעת הבחירה יש משמעות מסוימת, התשובה על כךח\.השימוש במשטח, בטעם האישי שלכם מבחינה עיצובית ועוד

לפי הסרט המוזכר במסמך 199,991, עושה רושם שהדאטאסט, או לפחות החלק הזה שלו, הוא מאמצע העשור הקודם.

שמירת דאטאסט מסודר למטרות אימון טוקנייזר

.התא הבא מכין את הדאטא עבור מנגנון האימון שנכיר בחלק הבא

```
CORPUS_DIR = './oscar_heb'
if not os.path.exists(CORPUS_DIR):
    os.mkdir(CORPUS_DIR)

text_data = []
file_count = 0

for sample in tqdm(sampled_ds):
    # remove newline characters from each sample as we need to use exclusively as seperators
    sample = sample['text'].replace('\n', '\s')
```

נבדוק את תקינות פורמט השמירה.

מהדאטה שאספנו BERT אימון טוקנייזר של

הטוקנייזר של BERT נקרא וורדפיס (WordPiece). האלגוריתם לאימון שלו דומה מאוד ל-BPE, ואופן חלוקת המילים לתמניות (token) ("פיסות" ("pieces")) שלו חמדני. בוורדפיס מסמנים פיסה שאינה בתחילת מילה על-ידי התחילית ##.

נאתחל טוקנייזר ריק ונאמן אותו על הדאטא שלנו. באתחול, צריך לפרט אילו פעולות טרום-טיקנוז (pre-tokenization) אנחנו רוצים לעשות, ובאימון צריך לפרט את ההיפר-פרמטרים של אוצר המילים המבוקש (נבחר ב-3,000), השכיחות המינימלית של טוקנים שיצורפו לאוצר המילים, ומגבלה על גודל האלפבית ההתחלתי. שני הפרמטרים האחרונים של train () נועדו להתאים את אוצר המילים שניצור לפורמט המקובל של הרובוטריק BERT.

האימון צפוי לקחת כ-10 דקות.

... ונטען אותו באמצעות האובייקט המתאים.

מי שמעונייןת מוזמןת לשמור את הקבצים בנפרד מהמחברת ולטעון לפי הצורך בשלבי פתרון התרגיל כדי לחסוך את זמן האימון עם כל סשן. עם זאת, ההגשה הסופית חייבת להיות תוצר של הרצה אחת נקייה מההתחלה ועד הסוף.

הבה נסתכל על כמה דוגמאות:

כמו שאנחנו רואים, הטוקנייזר שלנו בהחלט יודע לעבד עברית יותר טוב מאשר אנגלית, בהתאם לדאטא שקיבל לאימון.

סעיף 1 (0.5 נק'): הציעו הסבר מדוע בכל-זאת "שלום" מפריד את האות הסופית אך "עולם" מפריד שתי אותיות אחרונות. ניתן להוסיף תא קוד להנמקה.

הסבר אפשרי הינו ש'שלו' הינו נפוץ יחסית ולכן קיבל טוקן משל עצמו ,מאחר וזהו אלגוריתם גרידי הוא תחילית ארוכה יותר מ'של' ואילו 'עול' ככל הנראה לא נפוץ מספיק בשביל לקבל טוקן משל עצמו ולכן 'עו' קיבל את התחילית הארוכה ביותר (למטה ניתן לראות שעולים ועושים שניהם מתפרקים לטוקן 'עו')

כפי שאנו זוכרים מהשיעור, לרוב מאמנים טוקנייזר על אוצר מילים בסדר גודל של 30-50 אלף, ושלנו קטן פי עשרה. התוצאה היא הרבה פחות מילים וחלקי מילים שימושיים שזכו לקבל תמנית משלהן.

סעיף 2 (0.5) נק'): מצאו מילה (רצף אותיות עבריות ללא רווחים) תקינה בעברית שמתפצלת ליותר מ-6 תמניות.

```
### 3 ###

MY_WORD = 'האינצקלופדיה'

### # ###

toks = tokzr.tokenize(MY_WORD)

print(len(toks), toks)

7 ['האי', '##', '##', '##', '##', '##', '##']
```

כפי שהזכרנו בכיתה, אופן ה*פענוח* (decoding, inference) של אלגוריתם וורדפיס הינו חמדני (greedy): בהינתן רצף תווים, וורדפיס יחפש את התמנית הארוכה ביותר שמתאימה לתחילתו, וימשיך מהנקודה בה נקטע.

סעיף 3 (1 נק'): ממשו את אלגוריתם הפענוח באופן נכון אך לא בהכרח יעיל, באמצעות אובייקט הקורא את אוצר המילים ופועל חמדנית. הקוד בתא השני יבדוק את נכונותו.

```
from typing import Callable, List

class GreedyTokenizerFromFile(Callable):
```

```
def __init__(self, vocab_file: str) -> None:
   self.vocab = set()
   ### 5.1 (0.25 points) ###
   self.vocab |= set(open(vocab_file, 'r', encoding='utf-8').read().splitlines())
   print(f'initialized with {len(self.vocab)} pieces')
  def __call__(self, s: str) -> List:
   ### 5.2 (0.75 points, conditioned on test passing) ###
   tokens = []
   word = s
   while word != '##':
     for i in range(len(word), 0, -1):
       if word[:i] in self.vocab:
         tokens.append(word[:i])
         word = '##' + word[i:]
         break
       elif i == 1:
         tokens.append('[UNK]')
         return tokens
   return tokens
  ### you may define any additional functions below ###
### Do not modify this cell when submitting ###
from tokenizers.pre_tokenizers import BertPreTokenizer
greedy_tok = GreedyTokenizerFromFile(f'{MODEL_DIR}/wp-heb-vocab.txt')
# Bert's PreTokenizer handles punctuation. It's called implicitly by tokzr
pretok = BertPreTokenizer()
# our "test set" is a random document from the dataset
for w, char_span in tqdm(pretok.pre_tokenize_str(sampled_ds[2023]['text'])):
 # ground truth - the automatically-learned tokenization
 trained_tok = tokzr.tokenize(w)
 # preparation - BertTokenizer actually has a preprocessing submodule called
                 "basic_tokenizer" that lowercases English and such
 grd_tok = greedy_tok(tokzr.basic_tokenizer.tokenize(w)[0])
 # Test - these asserts should not fail #
 assert trained_tok == grd_tok, f'{w}, {trained_tok}, {grd_tok}'
    initialized with 3000 pieces
                   1 a/25 [aa·aa/?
                                  ?i+/cl
```

ימשחק השבת

נניח שאנחנו מתכננים משימה שנוגעת לעיבוד טקסט ממסמכים הנוגעים לספורט. בואו נראה עד כמה המודל שלנו מותאם למילים מתוך התחום (domain) הספורטיבי:

```
' '.join(tokzr.tokenize('מכבי הפועל כדורגל כדורסל כדורסל כדורעף טניס'))

→ "מכ ##ני ##ע ##ע ##ע ##ע ##ני ##ט ##ני ##סי
```

נראה שהמצב די רע. ננסה לאמן מודל טוקניזציה שיטפל במילים האלה יותר טוב **רק באמצעות שינוי קורפוס האימו**ן.

סעיף 4 (2 נק', אין ניקוד חלקי): צרו דאטאסט חדש מתוך הקורפוס שכבר בידינו (אל תייבאו קורפוס חדש, אל תפצלו מסמכים, ואל תכתבו יותר מ-200,000 מסמכים לדיסק) ואמנו עליו מודל WordPiece עם אותם היפר-פרמטרים כמקודם, שמפצל לכל היותר מילה אחת מהמילים לעיל ליותר משתי תמניות, ולא מפצל בכלל לפחות שתיים מהמילים. הראו את הפלט. תעדו היטב באמצעות הערות את כל השינויים שערכתםן ביחס לקוד המקורי.

צוותים שיפנו בשעות הקבלה שלפני תחילת השיעור ב-7/1 עם פתרון חלקי יוכלו לקבל רמז. לאחר מועד זה לא תינתן עזרה על הסעיף.

```
import re
def clean_text(text: str) -> str:
   Cleans the input text by removing characters that are not Hebrew letters, digits, spaces, or specified punctuation/symbols.
   Args:
       text (str): The input text to be cleaned.
   Returns:
      str: The cleaned text with only allowed characters and no extra spaces.
   cleaned_text = hebrew_digits_punct_regex.sub("", text)
   cleaned_text = re.sub(r'\s+', ' ', cleaned_text).strip()
   return cleaned_text
text = "45 ()*&^%$#@ (: ?האם זה עובד? ...) מבחן 123, האם זהו מבחן 231, האם זה עובד?
cleaned_text = clean_text(text)
print(cleaned_text)
→ 45 ()*&^%$#@ (: ?דבר האם זה מבחן 123, האם זה עובד? ישלום! זהו מבחן 123,
def download_corpus(keywords: List[str], n_docs: int) -> List[dict]:
   Downloads a corpus of documents containing specified words.
   This is the ssame as before, just filtering the documents by specified keywords.
   Args:
       words (List[str]): A list of words to search for in the documents.
       n_docs (int): The number of documents to download.
   Returns:
       List[dict]: A list of dictionaries, each representing a document that contains any of the specified words.
   i = 0
   sampled_sports_ds = []
   with tqdm(total=n_docs, desc="Finding sports documents") as pbar:
       for doc in dataset:
           if any([word in doc['text'] for word in keywords]):
               sampled_sports_ds.append(doc)
               i += 1
               pbar.update(1)
               if i >= n_docs:
                   break
   return sampled_sports_ds
import shutil
def save_corpus(corpus: List[dict], directory: str, clean=False) -> None:
   Save a cleaned corpus to a specified directory in chunks of 5000 samples.
   Args:
       corpus (List[dict]): A list of dictionaries, each containing a 'text' key with the text sample.
       directory (str): The directory where the cleaned text files will be saved.
    Returns:
       None
   if os.path.exists(directory):
       shutil.rmtree(directory)
   os.mkdir(directory)
    text_data = []
   file_count = 0
   for sample in tqdm(corpus, desc="Saving corpus"):
        # remove newline characters from each sample as we need to use exclusively as seperators
       sample = sample['text'].replace('\n', '\s')
       text_data.append(sample)
        if len(text_data) == 5_000:
           # once we hit the 5K mark, save to file
```

```
with open(f'{directory}/text_{file_count}.txt', 'w', encoding='utf-8') as fp:
                text = clean_text('\n'.join(text_data)) if clean else '\n'.join(text_data)
                fp.write(text)
            text_data = []
            file_count += 1
   \mbox{\tt\#} after saving in 5K chunks, we may have leftover samples, we save those now too
   with open(f'{directory}/text_{file_count}.txt', 'w', encoding='utf-8') as fp:
        text = clean_text('\n'.join(text_data)) if clean else '\n'.join(text_data)
        fp.write(text)
→ <>:24: SyntaxWarning: invalid escape sequence '\s'
     <>:24: SyntaxWarning: invalid escape sequence '\s'
     C:\Users\Benzu\AppData\Local\Temp\ipykernel_35672\3057535395.py:24: SyntaxWarning: invalid escape sequence '\s'
       sample = sample['text'].replace('\n', '\s')
SPORTS WORDS = 'מכבי הפועל כדורגל כדורסל כדורעף טניס'.split()
SPORTS_CORPUS_DIR = './sports_adjacent-clean'
SPORTS_MODEL_DIR = './sp-heb'
MODEL NAME = 'sports adjacent-clean'
N_DOCS = 50_000
### 4 ###
# download the corpus with the sports keywords
sampled_sports_ds = download_corpus(SPORTS_WORDS, N_DOCS)
# clean and save the corpus
save_corpus(sampled_sports_ds, SPORTS_CORPUS_DIR, clean=True)
# All of this is the same as before, but with the new corpus
sports_tokzr = tokzr_sports_trainer = BertWordPieceTokenizer(clean_text=True)
sports_paths = sorted([str(x) for x in Path(SPORTS_CORPUS_DIR).glob('**/*.txt')])
tokzr_sports_trainer.train(files=sports_paths, vocab_size=3_000, min_frequency=2,
                limit_alphabet=1_000, wordpieces_prefix='##',
                special tokens=[
                    '[PAD]', '[UNK]', '[CLS]', '[SEP]', '[MASK]'])
if not os.path.exists(MODEL_DIR):
 os.mkdir(MODEL DIR)
tokzr_sports_trainer.save_model(MODEL_DIR, f'{MODEL_NAME}')
Finding sports documents:
                                              | 0/50000 [00:00<?, ?it/s]
                                 0%
     Saving corpus: 0%|
                                   | 0/50000 [00:00<?, ?it/s]
     ['./wp-heb\\sports_adjacent-clean-vocab.txt']
sports_tokzr = BertTokenizer.from_pretrained(f'{MODEL_DIR}/{MODEL_NAME}-vocab.txt')
### # ###
' '.join(sports_tokzr.tokenize('מכבי הפועל כדורגל כדורסל כדורעף טניס'))
🚁 c:\Users\Benzu\anaconda3\Lib\site-packages\transformers\tokenization_utils_base.py:1925: FutureWarning: Calling BertTokenizer.from_pretr
       warnings.warn(
     מכבי הפועל כדורגל כדור ##סל כדור ##ע ##ף ט ##ניס'
    4
```