Mohammad Arabi 206985533
Saju Assi 314831207

דוו"ח
תרגיל בית 4

חלק א' :

בשלב הזה בהתחלה קראנו המשפטים מהקובץ .jsonl ואז אחרי זה עבור כל משפט פרקנו לטוקנים ורק מתוך הטוקנים של המשפט לקחנו הטוקנים שהם מילים בעברית ואז הכנסנו הטוקנים של המשפט הזה לרשימה שהרשימה הזאת גם הכנסנו אותה לרשימה בשם **tokenized_sentences** רשימה של רשימות הטוקנים של כל משפט. ואחרי זה עשינו סעיף ב ו-ג כמו מבוקש.

השאלות :

1) vector_size הוא גודל הווקטור שמייצג המילה במרחב הווקטורים כלומר כמה מימדים יהיה הווקטור שמייצג המילה.
- חסרונות הגדלת גודל הווקטור :
זמן האימון יהיה יותר גדול וגם זה יכול לגרום ל-overfitting הדבר שישפיע לרע.
- יתרונות הגדלת גודל הווקטור :
יהיה לנו יותר מידע כלומר יותר תכונות ואז זה יוביל ליותר דיוק בפעולות.
- חסרונות הקטנת גודל הווקטור :
יהיה לנו פחות תכונות ומידע הדבר שמקטין את הדיוק.
- יתרונות הקטנת גודל הווקטור :
זמן אימון יהיה קצר וגם פחות סיכוי ל-overfitting.

2) יש לנו למשל את הבעיה של מילים מחוברים למשל "בית משפט" או "בית חולים" כך ש משני מילים ביחד נקבל ביטוי הגיוני כי בית לבד או משפט לבד זה משהו אחר. בנוסף יש לנו בעיה של אותיות מחוברות עם מילים למשל: "ועד שהגיע" המילה "ועד" אפשר להבין אותה בשני דרכים שונים שלפי הרונטקסט אפשר לקבוע.

חלק ב' :

בשביל שלב א עברנו בלולאה עבור כל מילה ב-9 המילים שנתונים לנו ואז עבור כל מילה מהם בדקנו היחד בינה לבין כל מילה בקורפוס שלנו (כלומר המילים שבמודל שבינינו) בעזרת הפקודה similarity, ואז אחרי הבדיקות עבור כל מילה מ-9 המילים בחרנו ה-5 מילים הכי דומות לה במודל ואז כתבנו אותם בקובץ knesset_similar_words.txt .

בשביל שלב ב עברנו על משפט ברשימה **tokenized_sentences** ואז בכל פעם בהתחלה מגדירים ווקטור של 50 אפסים ואז אחרי זה נעבור עבור כל מילה במשפט הזה (רשימה של טוקנים) נבדוק אם היא במודל שלנו אם כן נוסיף להווקטור את **word_vectors[word]** (כאשר **word_vectors = model.wv**) ועלה הספירה ב-1 ואחרי כל החישוב נחלק הווקטור בקאונטר.

בשביל שלב ג אחרי שבחרנו 10 המשפטים השתמשנו בפונקציה **embeddings_of_sentences** שהגדרנו (שמקבלת רשימת משפטים ומודל) כדי לחשב ה-embedding של 10 המשפטים שבחרנו ואחרי זה חשבנו הדמיון בעזרת **cosine_similarity** בין ה-embedding של 10 המשפטים וה-embedding של המשפטים בדאתא שלנו.

בשביל שלב ד כתבנו 4 המשפטים כאשר המילה שבאדום תהיה עם [] (כלומר [מילה _אדומה])ואז בשימוש ב-**most_similar** נסינו עבורה הרבה נסיונות ומילים שונים בכל פעם שייתכן עבורם לקבל תוצאות מתאימות כלומר נסינו לבחור מילות שעבורם יתכן לקבל תוצאות מתאימות, ועבור כל מילה לקחנו מילה הכי מתאימה שקבלנו כך הגדרנו המילות האילו ידנית (הכי מתאימות שקבלנו עבור הפקודה **most_similar**) ושמנו במקום המילות באדום (עברנו על המשפטים ועבור המילות שמתחילות ב-[ומסתיימות ב-] שהם מילות אדומות הבדלנו אותם עם המילה המתאימה שמצאנו) כתבנו זה המילות החדשות ידנית כי הפקודה תתן תוצאות שונות. אילו הם התוצאות שרואים הם טובות ומתאימות שקבלנו עם מה בחרנו:

- משפט ראשון מילה אדומה לדיון בחרנו במקומה ב-למליאה שהיא השנייה מבין ה-3 הכי דומות :

```
[('שבמסגרת', 0.9472965002059937), ('למליאה', 0.9383843541145325), ('יתקיים', 0.9293624758720398)]
```

- משפט שני מילה אדומה מוכנה בחרנו במקומה ב-יכולה שהיא הראשונה מבין ה-3 הכי דומות :

```
[('יכולה', 0.818346381187439), ('חייב', 0.8069127798080444), ('לבתר', 0.80668985843658 45)]
```

- משפט שני מילה אדומה ההסכם בחרנו במקומה ב-התהליך שהיא הראשונה מבין ה-3 הכי דומות :

```
[('התהליך', 0.9397198557853699), ('הדוח', 0.9313275218009949), ('המהלך', 0.9290666580200195)]
```

- משפט שלישי מילה אדומה טוב בחרנו במקומה ב-בריא שהיא השנייה מבין ה-3 הכי דומות :

```
[('חכם', 0.921226978302002), ('בריא', 0.9138247966766357), ('בחור', 0.9116674065589905)]
```

- משפט שלישי מילה אדומה פותח בחרנו במקומה ב-עוצר שהיא השלישית מבין ה-3 הכי דומות :

```
[('סיימתי', 0.892031729221344), ('התייחסתי', 0.8826464414596558), ('עוצר', 0.8766990303993225)]
```

- משפט רביעי מילה אדומה שלום בחרנו במקומה ב-רבותיי שהיא השנייה מבין ה-3 הכי דומות :

```
[('הנוכחים', 0.9081310629844666), ('רבותיי', 0.9071613550186157), ('אוקיי', 0.9056839346885681)]
```

- משפט רביעי מילה אדומה להודיע בחרנו במקומה ב-להודיעכם שהיא הראשונה מבין ה-3 הכי דומות :

```
[('להודיעכם', 0.9309117197990417), ('הנני', 0.9263291954994202), ('מתכבדת', 0.8813149929046631)]
```

- משפט רביעי מילה אדומה היקר בחרנו במקומה ב-הנאמן שהיא הראשונה מבין ה-3 הכי דומות :

```
[('הנאמן', 0.9270415902137756), ('מהליכוד', 0.9242457151412964), ('לשאלות', 0.9224401116371155)]
```

השאלות :

1) כן יש כמה מילים שקבלנו שכן תואמות לציפיות שלנו למשל :
ישראל -> מדינתי

כנסת -> שזכויותיהם, מועצה

ועדה -> החלטה

ומילים דומות כמו למשל חבר שחבר שה כמובן יקבלו ערכים גדולים כי הם כמעט זיהים(יש הבדל באות).

ולפעמים יש מילים שבכלל לא קשורים אבל בכל מקרה קיבלו תוצאות גדולות 0.7+

כמו: ישראל ו למהותן שקיבלה 0.73.

אנחנו חושבים שהסיבה העיקרית היא גודל הקורפוס, אם היה לנו אחד יותר גדול היה יכולים להגיע לתוצאות יותר טוב ומילים יותר רלוונטיות.

2) מרחק קטן כי אם למשל ניקח משפט שיש בו מילה ונחליף אותה עם המילה ההפוכה אנחנו עדיין נקבל משפט הגיוני שאפשר להבין אותו למשל, להחליף את המילה "מותר" ב "אסור".

3) כמו שרואים מתוצאות שקיבלנו שה `antonyms` קיבלו תוצאות 0.55 ויותר ז"א שהמרחק ביניהן קטן כמו שהסברנו בסעיף קודם

```
0.9624152183532715אהבה ,שנאה:
0.925234854221344חם ,קר:
0.8305953145027161ראשון ,אחרון:
0.6510326266288757שמח ,עצוב:
0.5600370168685913 גדול ,קטן
```

4) אלה התוצאות שקיבלנו:

```
1  [. דבר אותו בדיוק זה :most similar sentence :. אבל זה אותו דבר
2   . את זה לתקן צריך פשוט :most similar sentence :. את זה בחשבון להביא צריך , ולכן
3   . את זה לבדוק מוכן אני , לא :most similar sentence :. אני לא מוכן לקבל את זה
4   . להצבעה עוברים אנחנו , רבותי , אם כן :most similar sentence :. להצבעה עוברים אנחנו ,רבותי , אם כן
5   . לא נורא שהוא דבר זה מבחינתך :most similar sentence :. זה לא דבר שהוא חדש
6   ? הזה בעניין תלסינקי אמנת אומרת מה :most similar sentence :? הזה בנושא שלכם התפקיד מה
7   . מדעתך שונה ודעתי , דעתי את אומר אני :most similar sentence :? את האמת אומר שאני בגלל
8   . הצבעה תהיה לא , אופן בכל , היום :most similar sentence :. היום תתקיים לא ההצבעה מקרה בכל
9   . בטוח כל כך לא אני :most similar sentence :. מבין כל כך לא אני
10  . עליו שחבל דבר , אתכם ללכת שצריך דבר זה :most similar sentence :? כזה דבר ייתכן איך
```

אני חושב שקיבלנו תוצאות טובות מאוד כך שרוב המשפטים קיבלנו באמת משהו דומה להן,למשל משפטים 1, 2, 4, 7, 3 אלה משפטים באמת דומים אחד לשני והן מהתוצאות הכי טובות שקיבלנו, אפילו יש משפטים שלא באופן ישיר דומות למשל משפט מספר 9 לא כל כך מבין ולא כל בטוח יכולות להיות אותו דבר (לא תמיד כמובן) אבל אם משיהו אומר שהוא לא מבין דבר שנאמר אז הוא לא בטוח על החלטה מסוימת או לא בטוח שהוא באמת הבין מה צריכים ממנו.

קיבלנו תוצאות כי לכל משפט המילים יש להם הסתברות גדולה להופיע ביחד ז"א שאפשר למצוא אותם כמה פעמים בקורפוס.

חלק ג' :

אחרי שקראנו המילים מקובץ `jsonl.` חלקנו המשפטים בשתי רשימות רשימה של committee ושנייה של plenary ואז בלולאה שבכל פעם לוקחת גודל `chunks` מבין שלוש הגדלים 1,3,5 כך בתחילת הלולאה נטעון את המודל דרך שימוש בפקודה Word2Vec.load שמקבלת נתיב למודל ואז נעשה כמו מה עשינו בתרגיל בית 3 שהיא חלוקה ל-chunks לפי הגודל הנוכחי ואז אחרי זה נעשה down_sample ואחרי נעשה טוקינציה כמו מה עשינו בחלק א' שלב א ואז על רשימת הטוקינציה נייצר sentence_embeddings בדיוק כמו שעשינו בחלק ב' סעיף ב כך sentence_embeddings זה כאילו ה-features שלנו, אחרי זה נאמן את המודל KNN ונעשה הפקודה train_test_split ואחרי זה KNN.fit ואז KNN.predict וא עכשיו נוכל לשתמש ב-classification_report. התוצאות שקבלנו :

- chunk_size = 1 :

	precision	recall	f1-score	support
committee	0.63	0.67	0.65	4298
plenary	0.65	0.61	0.63	4298
accuracy			0.64	8596
macro avg	0.64	0.64	0.64	8596
weighted avg	0.64	0.64	0.64	8596

- chunk_size = 3 :

	precision	recall	f1-score	support
committee	0.68	0.69	0.69	5731
plenary	0.69	0.68	0.68	5731
accuracy			0.68	11462
macro avg	0.68	0.68	0.68	11462
weighted avg	0.68	0.68	0.68	11462

- chunk_size = 5 :

	precision	recall	f1-score	support
committee	0.71	0.72	0.72	6591
plenary	0.72	0.71	0.71	6590
accuracy			0.71	13181
macro avg	0.71	0.71	0.71	13181
weighted avg	0.71	0.71	0.71	13181

השאלות :

1) התוצאות היו יותר גרועות, כך אילו התוצאות שקבלנו עכשיו (האחרון כאשר chunk size = 5):

```
              precision    recall  f1-score   support

   committee       0.63      0.67      0.65      4298
     plenary       0.65      0.61      0.63      4298

    accuracy                           0.64      8596
   macro avg       0.64      0.64      0.64      8596
weighted avg       0.64      0.64      0.64      8596

              precision    recall  f1-score   support

   committee       0.68      0.70      0.69      5731
     plenary       0.69      0.67      0.68      5731

    accuracy                           0.69     11462
   macro avg       0.69      0.69      0.69     11462
weighted avg       0.69      0.69      0.69     11462

              precision    recall  f1-score   support

   committee       0.72      0.73      0.72      6591
     plenary       0.72      0.71      0.72      6590

    accuracy                           0.72     13181
   macro avg       0.72      0.72      0.72     13181
weighted avg       0.72      0.72      0.72     13181
```

2) ההבדל הזה נובע לרוב בגלל שה feature vector שיש לנו הוא פחות מייצג מאשר ה feature vector שהיה לנו במשימה הקודמת.

3) אנחנו מקבלים שגודל צ'אנק 1 הוא גרוע יותר מגודל 3 ו 3 פחות טוב מ 5.(אז 5 היה הכי טוב מבין שלושה גדלים אלו) הסיבה היא שאם גודל הצ'אנק קטן, מטריצת התכונות תהיה גדולה יותר, מה שגורם לזמן הריצה לגדול עוד יותר. בנוסף, כל צ'אנק יכיל תכונות שלא מייצגות את הסוגים בצורה טובה מספיק, מכיוון שכוח features יתפזר בין הצ'אנקים הקטנים ולא יהיה מרוכז בצ'אנק אחד, מה שמקשה על משימת הסיווג.

חלק ד' :

אחרי שקראנו את המשפטים מהקובץ masked_sentences.txt הגדרנו
tokenizer = AutoTokenizer.from_pretrained('dicta-il/dictabert')
model = AutoModelForMaskedLM.from_pretrained('dicta-il/dictabert')
כפי ראינו באתר ואז עבור כל משפט בקובץ הנתון אחרי שהחלפנו כל [*] ב-[MASK] עברנו על כל [MASK] במשפט הנוכחי השתמשנו ב-model(tokenizer.encode שמשתמשת במודל כדי לבצע החיוי עבור המשפט ואז נשמור התוצאה מזה ונחשב האינדקס של ה-[MASK] ואז ב-torch.topk שנקבל ממנה הטוקן הסביר ביותר שיחליף [MASK] ואז נשאר להחליף המזהה לטוקן מילולי המתאים ואז בזה נקבל הטוקן שיחליף [MASK] אז נשמור אותו ברשומת הטוקנים וגם נחליף אותו עם ה-[MASK].

השאלות :

1) התוצאות הגיניות מאוד גם מבחינה תחבירית גם מבחינת משמעות המשפט.(התוצאות בתחילת העמוד הבא)

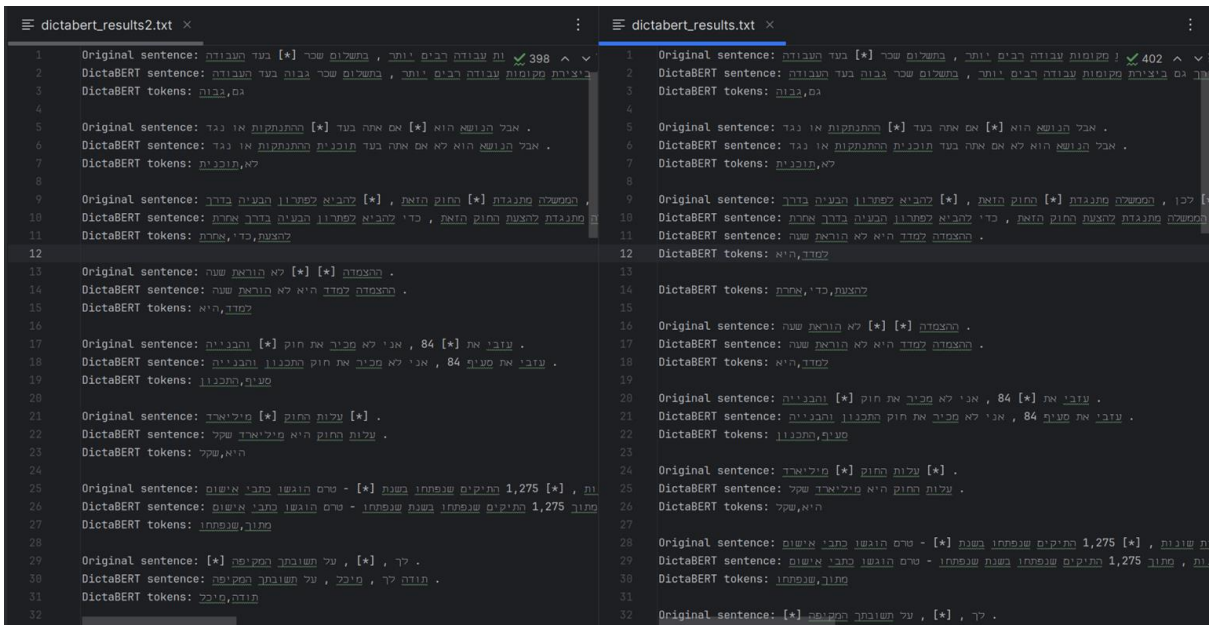
```
Original sentence: יש צורך [*] ביצירת מקומות עבודה רבים יותר , בתשלום שכר [*] בעד העבודה .
DictaBERT sentence: יש צורך גם ביצירת מקומות עבודה רבים יותר , בתשלום שכר גבוה בעד העבודה .
DictaBERT tokens: גם,גבוה

Original sentence: אבל הנושא הוא [*] אם אתה בעד [*] ההתנתקות או נגד .
DictaBERT sentence: אבל הנושא הוא לא אם אתה בעד תוכנית ההתנתקות או נגד .
DictaBERT tokens: לא,תוכנית

Original sentence: לכן , הממשלה מתנגדת [*] החוק הזאת [*] להביא לפתרון הבעיה בדרך [*] .
DictaBERT sentence: לכן , הממשלה מתנגדת להצעת החוק הזאת , כדי להביא לפתרון הבעיה בדרך אחרת .
DictaBERT tokens: אחרת,כדי,להצעת

Original sentence: [*] [*] לא הוראת שעה ההצמדה .
DictaBERT sentence: ההצמדה למדד היא לא הוראת שעה .
DictaBERT tokens: היא,למדד

Original sentence: עזבי את [*] 84 , אני לא מכיר את חוק [*] והבנייה .
DictaBERT sentence: עזבי את סעיף 84 , אני לא מכיר את חוק התכנון והבנייה .
DictaBERT tokens: סעיף,התכנון

Original sentence: מיליארד [*] עלות החוק [*] .
DictaBERT sentence: מיליארד שקל עלות החוק היא .
DictaBERT tokens: שקל,היא

Original sentence: מסיבות שונות , [*] 1,275 התיקים שנפתחו בשנת [*] - טרם הוגשו כתבי אישום .
DictaBERT sentence: מסיבות שונות , מתוך 1,275 התיקים שנפתחו בשנת שנפתחו - טרם הוגשו כתבי אישום .
DictaBERT tokens: מתוך,שנפתחו
```

אבל היה לנו גם תוצאות שהן פחות טובות ופחות הגיוניות:

```
Original sentence: מסיבות שונות , [*] 1,275 התיקים שנפתחו בשנת [*] - טרם הוגשו כתבי אישום .
DictaBERT sentence: מסיבות שונות , מתוך 1,275 התיקים שנפתחו בשנת שנפתחו - טרם הוגשו כתבי אישום .
DictaBERT tokens: מתוך,שנפתחו
```

2) כן כמו רואים היה שיפור בתוצאות כך עכשיו לרוב קבלנו תוצאות שיותר מתאימות ממה קבלנו בתרגיל בית 2 אבל גם יש משפטים שבהם בתרגיל בית 2 קבלנו תוצאה יותר טובה אבל זה במעט מהמקרים, אז בכללי לרוב כן קבלנו שיפור בתוצאות.

3) כן נקבל אותו הדבר:



זה משום ש מודל BERT אינו מודל רנדומלי ז"א אותו קלט עם אותם פרמטריים תמיד יניבו לנו תוצאות דומות וזה באמת מה שמקבלים.

4) כן כמו שהראינו בסעיף 1:

```
Original sentence: מסיבות שונות , [*] 1,275 התיקים שנפתחו בשנת [*] - טרם הוגשו כתבי אישום .
DictaBERT sentence: מסיבות שונות , מתוך 1,275 התיקים שנפתחו בשנת שנפתחו - טרם הוגשו כתבי אישום .
DictaBERT tokens: מתוך,שנפתחו
```

למשל המשפט הזה, אז כנראה המודל הוא פחות טוב בחיזוי מספרים כמו שנה חודש או יום זה יכול גם להיות כי זה יכול איזה שנה והוא לא יכול איזה בדיוק לכן גם אם נתן מספר אולי יהיה לא נכון כי יש אינסוף אפשריות.(זה למרות שחיזה טוקן תקין אבל הוא לא נכון).

```
Original sentence: בסופו של דבר , גם אני הייתי [*] אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף [*] לא יודע שבאמת אנחנו מייצגים את ישראל .
DictaBERT sentence: בסופו של דבר , גם אני הייתי בשתי אליפויות העולם רצוף שבע שנים , וכן יצאנו עם דגל ישראל ואני יודעת שחוץ מאימא שלי , אף אחד לא יודע שבאמת אנחנו מייצגים את ישראל " .
DictaBERT tokens: אחד,בשתי,"
```

בדוגמה זו היה לני גם את ' " ' שהיה בלי הסוגר שלו. שזה לא נכון מבחינת כתיבה.