דו"ח תרגיל בית 4

Word Embeddings

שמות המגישים + ת"ז:

עוביידה חטיב, 201278066

מאיה עטואן, 314813494

חלק א'

א. המשפטים חולצו מקובץ ה- JSONL, ועברו טוקנזציה תוך כדי הסרת טוקנים שאינם מלים. כל משפט נשמר כרשימה של טוקנים של המלים המרכיבות אותו, שהתווספה לרשימה של רשימות.

הטוקן נחשב למילה (ועל כן נהיה חלק מהטוקנים שמרכיבים את הרשימה של המשפט) אם הוא מקיים את ארבעת התנאים הבאים:

1. הוא **מתחיל באות עברי**.
2. הוא **מסתיים באות עברי** או **בגרש (')**.
3. **האותיות שבאמצע (חוץ מהראשונה והאחרונה) מורכבות, מאותיות עבריות, מגרש (') או מגרשיים (")**.
4. **הוא מאורך לפחות 2,** אלא אם האות האחרונה היא ', אז האורך לפחות 3**.**

התנאי השני לקח בחשבון גם מלים שמסתיימות בגרש כמו את השם הפרטי סמית'. התנאי השלישי התחשב במלים שמכילות גרש או גרשיים בתוכם דוגמת ג'קוזי, ג'ינס או קיצורים של שמות ערים כמו ת"א ו- ב"ש שראינו לנכון להכיל אותם. התנאי הרביעי נועד להוציא טוקנים שמורכבות מאות אחת או אות עם גרש, טוקנים כאלה נפוצים בקורפוס שלנו בשל ההקשר שלו, ולא נחשבות למלים תקינות.

כמו כן, נלקחו בחשבון מקרים בהם המילה באה בין גרשיים, ע"י כך שאם המילה לא עונה על אחד משלושת התנאים הנ"ל אך מתחילה באחד מבין [',"], ומסתיימת בסימן זהה, והמילה שבאמצע (לא כולל האותיות הראשונה והאחרונה) עונה לארבעת התנאים שלמעלה, גם כן היא נחשבת למילה ורק החלק שבאמצע מתווסף כטוקן לרשימה. דוגמה לכך: "שיקול", המילה מתחילה ב- ["], ומסתיימת באותו סימן, והחלק האמצעי (שיקול) עונה על 4 הדרישות הנ"ל, לכן המחרוזת שנשמרת ברשימה עבור הטוקן הזה היא (שיקול).

ב. הארגומנט **window** מציין את כמות הטוקנים שנלקחים בחשבון לפני ואחרי הטוקן שנרצה לבנות לו וקטור. הערך של window נבחר להיות 5 מפני שהוא נותן הקשר טוב, ערך יותר קטן מפספס מלים שהם כן רלוונטיים לטוקן, וערך יותר גדול נוטה הרבה פעמים לכלול מלים שלא כל כך קשורות למשמעות של הטוקן או להקשר שלו.
הארגומנט **min_count** מציין את כמות ההופעות המינימלית של הטוקן בקורפוס על מנת להיכלל במטריצה. מפני שהקורפוס שלנו נחשב לקטן יחסית, החלטנו לקבוע את הערך של הפרמטר להיות 1 על מנת לכלול כמה שיותר מלים.
הארגומנט **vector_size** המציין את אורך הווקטור של הטוקן נבחר להיות 100, כך שהמספר מאזן בין גודל ייצוג מספיק טוב וסיבוכיות לא גבוהה.
הערכים נבחרו על סמך הביצועים שלהם על שתי המשימות של דמיון בין מלים ודמיון בין משפטים.

שאלות

1. הסבירו מה המשמעות, ומה היתרונות והחסרונות של הגדלת והקטנת גודל הוקטור- vector_size.

- וקטור גדול נותן ייצוג יותר טוב וביטוי יותר טוב לתכונות, מה שיכול להשתקף בדיוק יותר גבוה, לעומת וקטור קצר.
- שימוש בווקטור גדול עלול לייצור overfitting יותר מאשר שימוש בווקטור קטן.
- ווקטור גדול מצריך סיבוכיות זמן גבוהה יותר לעומת ווקטור קטן.
- אחסון ווקטורים גדולים מצריך יותר מקום לעומת ווקטור קטן.

2. **הסבירו מה הבעיות שיכולות לעלות משימוש במודל הנ"ל , שאומן על הקורפוס שלנו. התייחסו בתשובתכם לאופן שבו יצרנו את הקורפוס, לגודל שלו ולשימושים פוטנציאלים של המודל.**

- הקורפוס שלנו נחשב לקטן יחסית, ויכול להיות שהרבה מלים לא מיוצגות בו או מיוצגות אך לא במספיק הקשרים בהם המילה יכולה לבוא.
- הקורפוס שלנו לקוח מפרוטוקולי הכנסת, ויכול להיות שהביצועים שלו לא יהיו טובים מספיק על משפטים שנלקחו מהקשרים אחרים. סיבה אחת לכך יכולה להיות למשל בגלל שבדיוני הכנסת יש הקשרים אשר יותר נפוצים מאחרים באופן הלא מייצג את המציאות.
- בשלב היצירה של הקורפוס לא נעשתה חלוקה למורפמות בתהליך הטוקנזציה. דבר זה יכול להיות בעל הרבה השלכות שליליות. אחת מהם למשל היא שהמודל אף פעם

לא יחזה את המילה "וחבר" אם היא לא הופיעה בכלל בקורפוס למרות ש- "חבר" הופיעה מספיק פעמים. במקום זאת, הוא יחזה "חבר" או מילה אחרת שהופיעה מספיק פעמים עם האות "ו" לפניה.

- בחירת יחידת הסיווג כמשפט אחד, לצד היות הרבה משפטים בעלות אורך קצר, יכול לגרום לרעש.

חלק ב'

א. המשימה בוצעה בקוד שלנו ע"י הפונקציה 'most_similar_words' אשר מחשבת את הדמיון בין כל מילה ושאר המלים על סמך המודל שיצרנו, מסדרת אותם ומדפיסה לקובץ 'knesset_similar_words.txt' את 5 הראשונות ביחד עם ציון הדמיון בין שתי המלים.

ב. הנדרש מבוצע בקוד שלנו ע"י הפונקציה 'sentences_embed', אשר מקבלת כקלט את המשפטים כרשימה של מלים בלבד (שהתקבלו לאחר ניקוי של הטוקנים שאינם מלים), ואת ווקטורי המלים של המודל שחושבו קודם לכן. עבור כל משפט, הפונקציה בונה ווקטור שהוא ממוצע של ווקטורי המלים המרכיבות את המשפט. במידה ולמלים של משפט אין ייצוג כלשהו, הוא מיוצג ע"י ווקטור של אפסים. הפונקציה מחזירה רשימה של הווקטורים שיצרה עבור כל משפט.

ג. המשפטים נבחרו כך שיכילו מגוון רב של תכונות: חלקם לקוחים מפרוטוקולי והחלק האחר מפרוטוקולי המליאות, בעלי אורכים שונים, עוסקים בנושאים שונים, מורכבים רק מטוקנים שכיחים או שמכילים גם טוקנים נדירים, בעלי אופי חיובי, שלילי וניטרליים, נאמרים בגופי זכר, נקבה או רבים, מכילים שמות פרטיים ושמות מדינות או שלא.
עבור כל אחת מהמשפטים, ווקטורי Embedding מיוצרים באמצעות אותה פונקציה ' sentences_embed' מהסעיף הקודם.
הפונקצה most_similar_sentence מקבלת כקלט את המשפטים שנבחרו, ה- Embedding שלהם, ביחד עם שורות הקורפוס ו- ווקטורי ה- Embedding עבור כל משפט שבשורה, ובאמצעות מיתודת 'cosine_similarity' מוצאות עבור כל אחד מ- 10 המשפטים שנבחרו את המשפט הכי קרוב מתוך משפטי הקורפוס. לאחר מכן, היא מדפיסה את הפלט בפורמט הנדרש לקובץ 'knesset_similar_sentences.txt '.

ד. המשימה בוצעה ע"י הפונקציה 'tokens_replace_to_similar ' שנקראה מה- ()main וקבילה כפרמטרים את הווקטורים של המודל, את המשפטים ביחד עם הטוקנים שנרצה

להחליף וערכי פרמטרי ה- positive וה- negative של כל טוקן, ואת הכותרת של קובץ ה- txt שנרצה לשמור אליו את ההדפסה. הפונקציה עברה על כל הטוקנים שנרצה להחליף, ובאמצעות המיתודה "most_similar" עם ערכי ה- positive וה- negative של הטוקן ו- topn=1, מצאה את המילה הכי קרובה, החליפה אותה במשפט והדפיסה את שני המשפטים, הממוסך והחדש, ביחד עם המלים שהוחלפו.

בהתחלה נסינו את המשימה בלי להתייחס לפרמטרים positive, negative, והתוצאות שהתקבלו ברובם היו פגומות. כמה משפטים הוחלפו במלים הנכונות מבחינת משמעות אך לא כתחביר, למשל, "בעוד מספר דקות ..." הפכה להיות "בעוד מספר דקה ...". משפטים אחרים היו נכונים תחבירית אך שונים מבחינת משמעות, כמו " בוקר טוב , אני פותח את הישיבה ." שהפכה להיות " תקשיב טוב , אני עוצר את הישיבה .". משפטים אחרים לא היו נכונים לא כמשמעות ולא כתחביר כמו " בתור יושבת ראש הוועדה , אני מוכנה להאריך" שהפכה להיות " בתור יושבת ראש הוועדה , ואני מוכנה להאריך".

התמודדנו עם הבעיה דרך הוספת מלים לרשימה של ה- positive וה- negative, אך לרוב ל- positive. המלים שהוספנו אותם היו נרדפים או מלים מאותו הקשר למילה שנרצה להחליף במטרה לחדד את המשמעות של המילה במיוחד כאשר היא מקבלת יותר ממשמעות אחת. דוגמה לכך היא המילה "שלום" שמקבלת יותר משמעות של "קיום יחסים דיפלומטיים וסיום מצב מלחמה" בשל היות הקורפוס שלנו לקוח מדיונים בעלי הקשר פוליטי, אולם המשמעות שלה שונה לגמרי במשפט " שלום , אנחנו שמחים להודיע שחברינו היקר קיבל קידום". דרך אחרת לחידוד משמעות הייתה דווקא בהוספת ההפך של המילה ל- positive, זה קרה במילה "פותח" כאשר הוספנו ל- positive שלה את המילה "עוצר", ההיגיון מאחורי זה הוא שהרבה פעמים ההפך של המילה בא באותו הקשר גם כן. במקרים שבהם המילה שהתקבלה לא תואמת לצורת ריבוי המילה המוחלפת, התמודדנו עם כך ע"י להוסיף ל- positive מלים מאותה צורת ריבוי, ול- negative מילות שלא מאותה צורת ריבוי ובמיוחד מצורת הריבוי המתקבלות, זה קרה למשל במילה בוקר במשפט "בוקר טוב , אני פותח את הישיבה" שהייתה מוחלפת ב- "שבועיים" והתגברנו על כך ע"י להוסיף "ערב" ו- "חודש" ל- positive, ו- "שעתיים" ל- negative, והמשפט הפך להיות כתוצאה מכך: "שבוע טוב, אני אפתח את הישיבה".

לאחר התהליך המתואר למעלה, התוצאות נהיו יותר טובות מבחינה תחבירית, והמשפטים שהתקבלו היו יותר קרובות למשמעויות של המשפטים המקוריים.

שאלות

1. **האם המילים הכי קרובות שקיבלתם בסעיף א' תואמות את הציפיות שלכם? הסבירו. גם אם תאמו לציפיות וגם אם לא, נסו להסביר מדוע זה עבד או לא עבד טוב.**

רוב המלים שקיבלנו היו צפויות, אך חלק גדול לא. המלים הכי קרובות שהיו צפויות היו בעיקר משני סוגים: אלה הזהות למלים הנתונות אך בניתוח תחבירי שונה (למשל, 'וחבר', ו- 'לחבר' כדומות ל- 'חבר'), או מלים מאותו הקשר (כמו 'ממשלתנו' ו- 'עיריית' כדומות ל- 'רשות'). תוצאה מעניינת הייתה למלים הדומות ל- "גברת" שהיו ברובם שמות פרטיים של נקבות. המלים הדומות למילה "בוקר" (כמו "התקשר" ו- "שתשקול") לא היו צפויים, ולא נראה שיש להם קשר חזק אליה. אחד ההסברים לכך יכול להיות שהמילה מופיעה לרוב בתחילת המשפט, ולכן החלון שלה קטן ומורכב רק מ- 5 המלים הבאים אחריה. סיבה אחרת לקבלת מלים לא צפויות היא כאשר המילה שמחפשים דומות לה מופיעה בהקשרים מרובים, כך שאין לה הקשר ספציפי ומילות טיפוסיות שמתלוות אליה. סיבה נוספת שיש לה השפעה היא היות הקורפוס קטן ולא מכיל הקשרים רבים, והקשר מסוים לא טיפוסי (רעש) יכול לקבל משקל גדול.

2. **אם ניקח שתי מילים שנחשבות להפכים (antonyms), למשל "אהבה" ו"שנאה", או "קל" ו"כבד". האם היינו מצפים שהמרחק בין שני וקטורי המילים שלהן יהיה קצר או ארוך? הסבירו.**

נצפה שהמרחק בין שני הווקטורים של מילה וההופכית שלה יהיה קצר, משום ששתי המלים נוטות להופיע בהקשרים דומים. למשל, 'אהבה' ו- 'שנאה' מופיעות יותר במשפטים המדברים על רגשות.

3. **מצאו שלושה זוגות של מילים שנחשבות להפכים (antonyms) הקיימות בקורפוס שלנו ובדקו את המרחק ביניהן. האם הציפייה שלכם מסעיף 2 מתקיימת עבורן עם המודל שבניתם?**

להלן התוצאות שקיבלנו עבור 3 הזוגות: (מהיר, איטי), (ימין, שמאל), ו- (כבד, קל).
התוצאות תואמות במידה רבה לציפיות שלנו בסעיף 2.

```
0.7251496  :מהיר, איטי
0.92627686  :ימין, שמאל
0.8530906  :כבד, קל
```

4. **האם המשפטים הכי קרובים בסעיף ג' תאמו לציפיות שלכם? הסבירו. גם אם תאמו לציפיות וגם אם לא, נסו להסביר מדוע זה עבד או לא עבד טוב.**

המשפטים שקיבלנו תאמו ברובם לציפיות שלנו של לקבל משפטים דומים כאלה הדומות בהקשר או בנושא שבו הן דנות או באופי החיובי\השלילי שלהן. במשפט אחד (האחרון בקובץ) התקבל משפט שדומה בדיוק למשפט שניתן כקלט, וכשחזרנו לקורפוס התברר שהמשפט הופיע פעמיים ושזאת ההופעה האחרת שלו. ההצלחה של הקורפוס נובעת

מכך שהצליח בלמידת הקשר הסמנטי של המלים. מפני שמלים נוטות לבוא בהקשרים מסוימים ולהתלוות אחת לשנייה, האלגוריתם נתן למשפטים המכילים אותן את אותם ווקטורים שקרובים בערכים שלהם. ניתן לראות שהאלגוריתם הצליח במיוחד עם (1) משפטים שאפשר לזהות בהם נושא ספציפי יותר מאשר משפטים כלליות, כנראה בשל כך שהן מכילות יותר קולקציות הנוטות להופיע באותם הקשרים ספציפיים, (2) משפטים ארוכים יותר ממשפטים קצרים, סביר להניח שבגלל היותם מכילים יותר הקשרים מה שגורם ליותר דיוק.

חלק ג'

פיתחנו את הקוד עבור המשימה בהתבסס על אותם פונקציות של המשימה הבינארית מתרגיל בית 3, עם השינוי שבמקום ווקטורי המאפיינים הקוד כעת ישתמש ב- sentence embedding של משפטי שני הדוברים. ה- embedding של המשפטים נעשה באותה דרך שהוסברה בחלק ב', סעיף ב', ובהתבסס על המודל שבנינו מקודם. הערכים של הפרמטרים של מודל ה- KNN היו דומים לערכים של הפרמטרים שלו מתרגיל בית 3, שהם כערכי ברירת המחדל, מלבד n_neighbors=8, ו- weights='distance'. הסיבה לשימוש באותם ערכים של הפרמטרים הוא על מנת לעשות השוואה הוגנת. התוצאות שהתקבלו מצורפות להלן:

	precision	recall	f1-score	support
speaker1	0.77	0.86	0.81	2159
speaker2	0.84	0.74	0.79	2159
accuracy			0.80	4318
macro avg	0.80	0.80	0.80	4318
weighted avg	0.80	0.80	0.80	4318

1. **האם עבור אותם פרמטרים ותנאים שהשתמשתם בהם בתרגיל 3 קיבלתם תוצאות טובות יותר או פחות עבור וקטור המאפיינים הנ"ל? הסבירו מדוע לדעתכם זה קרה.**

התוצאות שהתקבלו היו פחות טובות בהשוואה לתוצאות שהתקבלו עבור שני ווקטורי המאפיינים (וקטור ה- TF-IDF והווקטור המותאם אישית שבנינו בתרגיל הבית הקודם). הסיבה לכך יכולה להיות שהווקטור של המשפט כעת לא משקף תכונות אישיות של הדובר כמו במקרה שני הווקטורים מהתרגיל בית הקודם, אלא יותר את המלים המרכיבות אותו וההקשרים שבהם הם מופיעים כך שמאפיינים כמו אורך המשפט, מספר

הכנסת וסוג הפרוטוקול לא מיוצגים בו. סיבה נוספת היא הטוקינזציה של המשפטים שעשינו בתחילת התרגיל שהשאירה רק את הטוקנים שהם מלים, מה שפגע בכמה מאפיינים שעליהם הסתמכנו בזיהוי הדובר בתרגיל הבית הקודם דוגמת אם המשפט מכיל ספרה או שלא. הסיבות שהוזכרו גורמות לווקטורים שיצרנו להיות פחות מתאים למשימה של הבחנה בין שני דוברים.

חלק ד'

הקוד מומש כך שהוא קורא את שני הקבצים: האחד עם המשפטים המקוריים, והשני עם המשפטים הממוסכים. המופעים של [*] מוחלפים ב- [MASK]. האלגוריתם, לאחר מכן, חוזה את הטוקנים החסרים באמצעות מודל DictaBERT, ולבסוף שומר את התוצאות לקובץ ' dictabert_results.txt' לפי המבנה הנדרש.

1. **האם קיבלתם משפטים הגיוניים? מבחינת התוכן, קוהרנטיות ומבחינה תחבירית.**
המשפטים שקיבלנו לרוב בעלי משמעות ונכונים מבחינה תחבירית.

2. **האם קיבלתם השלמות קרובות למילים החסרות האמיתיות? פרטו.**
חלק גדול מההשלמות היו של הטוקנים הנכונים, החלק האחר לא נחזו במדויק, אך במקום נחזו טוקנים הגיוניים שיצרו משפטים בעלי משמעות, שלפעמים תואמת למשמעות המקורית ולפעמים שונה. ניתן לשים לב, שהמלים שלא נחזו נכון לרוב נמצאות במשפטים ארוכים, ושהן נכונות להקשר הצר סביב המילה אך לא למשפט השלם.
דוגמה לטוקן שנחזה בצורה נכונה:
אני יכול להבטיח לך רק דבר אחד – שהפחד היחידי ...
אני יכול להבטיח לך רק דבר **אחד** – שהפחד היחידי ...

דוגמה למשפט שנחזה בצורה לא נכונה אך הטוקן החדש לא שינה את המשמעות של המשפט:
הראשונה היא אזרחי המדינה שתפסה את השטח
הראשונה היא אזרחי **ישראל** שתפסה את השטח

דוגמה למשפט שנחזה בצורה לא נכונה כך שהטוקן החדש שינה את משמעות המשפט אך שמר על היותו נכון תחבירית:

לכן אנחנו חושבים שנכון לתקוף את זה משתי הזוויות האלה

לכן אנחנו חושבים שנכון **לבחון** את זה משתי הזוויות האלה

3. **השוו את התוצאות שקיבלתם עכשיו לאלו שקיבלתם בתרגיל בית 2. האם יש שיפור בתוצאות לדעתכם?**

התוצאות שקיבלנו בעלות אחוז דיוק גבוה באופן משמעותי מהתוצאות של תרגיל בית 2. כמו כן, המלים המחליפות במשימה הנוכחית יוצרות משפטים יותר קוהרנטיים ונכונים תחבירית.

4. **האם יש משפטים שעבורם המודל עבד פחות טוב? אם כן ואם לא, הסבירו מה לדעתכם הסיבה לכך.**

לפי המדגם שלנו, ניתן לראות שהמודל עובד יותר טוב עבור המשפטים בעלי אורך קצר יותר מאשר משפטים ארוכים. הסיבה לכך לדעתנו היא שההקשר של המילה במשפטים הקצרים יהיה המשפט כולו (החלון לפני המילה ואחרי מכסה את כל המלים במשפט) ולכן המילה מותאמת יותר טוב למשפט. בנוסף, ניתן לשים לב שהמודל לא מצליח לחזות באופן מדויק את הטוקנים הממוסכים כאשר ההפרש ביניהם לא גדול (למשל, כאשר רק 1-2 טוקנים מפרידים בין הטוקנים הממוסכים), וזה יכול לנבוע מכך שהשלמה אחת שהיא לא בהכרח מדויקות משפיעה על ההשלמה שאחריה וכך הלאה.

5. **באילו מקרים לדעתכם יש למודל סיכוי גבוה יותר להיכשל או להחזיר תוצאה פחות טובה? מדוע?**

כפי שהוזכר בסעיף הקודם, אנחנו מצפים שהמודל יעבוד טוב עם משפטים קצרים, ומשפטים שהטוקנים הממוסכות בהם רחוקות אחת מהשנייה ואין חיתוך בין החלונות סביבם. כמו כן, אנחנו מצפים שהמודל יודע לחזות טוקנים יותר נפוצים כך שהוא תופס את ההקשרים שלהם בצורה יותר רחבה ומדויקת. כמו כן, שמנו לב שהמודל מתקשה עם חיזוי מילות שלא אופייניות להקשר ספציפי דוגמת המילה "לתקוף" במשפט " לכן אנחנו חושבים שנכון לתקוף את זה משתי הזוויות האלה".