NLP – HW4

מגישים:

מוחמד חג'אזי – 213011026

מועאד עבד אללטיף –209418896

חלק א׳: יצירה של מודל Word2Vec ואימונו על קורפוס הכנסת

ענו על השאלות הבאות:

1. היכנסו לתיעוד המודל, הבינו מה המשמעות של כל ארגומנט שנתנו לו (vector_size, window, min_count) והסבירו מה היתרונות והחסרונות של הערכים שנבחרו עבור השימוש שלנו. האם הייתם בוחרים ערכים אחרים ?

vector_size (100): הפרמטר הזה מגדיר את גודל הווקטורים של המילים במודל. גודל ווקטור של 100 אומר שכל מילה מיוצגת על ידי ווקטור בעל 100 ממדים, מה שיוצר איזון בין לכידת מידע סמנטי מספיק לבין יעילות חישובית. ווקטורים גדולים יותר יכולים ללכוד משמעויות עדינות יותר אך דורשים יותר זיכרון ומשאבי חישוב, בעוד שגדלים קטנים עלולים לפספס פרטים מסוימים. עבור הקורפוס שלנו שהוא דיאלוג, גודל של 100 מספק איזון טוב בין היכולת לתפוס את המשמעויות העשירות והדקיות הספציפיות לדיאלוג לבין שמירה על יעילות חישובית. הוא מאפשר למודל להבין ולהפריד בין דפוסי שיחה שונים תוך שמירה על מידת דיוק גבוהה ללא צורך במשאבים מוגזמים.

window (5): הפרמטר הזה מגדיר את המרחק המקסימלי בין המילה הנוכחית למילה שמנובאת בתוך המשפט, כאשר הוא מוגדר כאן ל-5. הגדרה זו אומרת שהמודל בוחן חמש מילים לפני וחמש מילים אחרי המילה המטרה לצורך הקונטקסט. גודל חלון של 5 יעיל ללכידת הקונטקסט המקומי הרלוונטי מבלי להכניס רעש ממילים לא קשורות. עם זאת, חלון גדול יותר עשוי לתפוס מידע רלוונטי יותר. הבחירה ב-5 היא אופטימלית לדיאלוגים, ומאזנת בין לכידת קונטקסט להפחתת רעש.

min_count (1): הפרמטר הזה מגדיר את המספר המינימלי של פעמים שמילה צריכה להופיע בקורפוס כדי להיכלל במודל. הגדרתו ל-1 כוללת את כל המילים, אפילו אלו שמופיעות רק פעם אחת (בת"ב 1 ראינו שיש המון כאלה), מה שעשוי להכניס רעש ממילים נדירות אך מבטיח שאף מילה פוטנציאלית רלוונטית לא תושמט. עבור קורפוסים גדולים יותר, הגברת הסף הזה עשויה לשפר את איכות המודל על ידי התמקדות במילים רלוונטיות ושכיחות יותר והתעלמות מחריגים. סף גבוה יותר, כמו 5 או 10, עשוי להועיל לשיפור איכות המודל על ידי הפחתת רעש והפחתת עומס חישובי, אך עלול להסתיר מידע ממילים פחות שכיחות.

אם ראינו שהמודל מלא רעש, אז שיקול min_count של 2 או 3 יכול לעזור להפחית רעש ולמקד את המודל במילים שבאמת משמעותיות להבנת הקונטקסט, בלי להסתכן באובדן מידע חשוב מדי.

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

אופן יצירת הקורפוס: מכיוון שהקורפוס נוצר מדיבורים בכנסת, הוא עשוי לשקף את השפה הפוליטית והפורמלית יותר מאשר שפה יומיומית או תחומית. זה עלול להגביל את השימושיות של המודל ליישומים ספציפיים ולא לתת תמונה מדויקת של השפה העברית בכללותה.

גודל הקורפוס: גודל הקורפוס יכול להשפיע על איכות המודל. קורפוס קטן עלול להביא למודל שאינו מסוגל ללמוד דפוסים מורכבים בשפה. גודל הקורפוס שלנו שמשמש את המודל הוא כ- 14,000 משפטים גודל זה נחשב קטן עד בינוני. גודל זה יכול להוות בסיס טוב ללמידת דפוסים בסיסיים בשפה, אך עשוי להיות מוגבל ביכולת לתפוס דקויות מורכבות ונדירות בלי שימוש בטכניקות מתקדמות או הרחבת הקורפוס.

התעלמות ממילים נדירות: על אף שהשימוש ב-min_count=1 מאפשר כלול את כל המילים, זה גם עלול לכלול רעש ולא לתת משקל מספיק למילים נדירות אשר יכולות להיות חשובות במיוחד בתחומים ספציפיים או בהקשרים מסוימים.

התמקדות בקונטקסט מילולי בלבד: המודל עוסק בלמידה של וקטורים למילים בהתבסס על הקונטקסט הסמוך, אך אינו מתחשב במרכיבים סימנטיים או סינטקטיים מורכבים אחרים של השפה. זה עלול להגביל את יכולתו להבין משמעויות מורכבות ודקויות שפתיות.

שימושים פוטנציאליים: השימושיות של המודל תלויה במטרות הספציפיות שלו. עבור ניתוח תוכן פוליטי, פורמלי או משפטי, המודל עשוי להיות מאוד מועיל. אולם, בשימושים שדורשים הבנה של שפה יומיומית, סלנג או טקסטים מתחומים אחרים, המודל עשוי לא להיות מדויק או אפקטיבי במיוחד. המודל יכול להתקשות לתפוס דקויות שפתיות וטקסטואליות שאינן מופיעות בתדירות גבוהה בדיבורים הפוליטיים והפורמליים.

חלק ב׳ : דמיון בין מילים

ד.red_words_sentences

.הסבירו בדו"ח מה ניסיתם, מה עשיתם,מה קיבלתם והאם הצלחתם במשימה:

במשפט הראשון רמינו קצת ונתננו שתי מילים כדי לקוון את המודל שיצוא מילה ספציפית :

```
similar_word = model.wv.most_similar(positive=['כפיים','למחוא'],negative=[],
topn=1)
```

קבלנו : לאולם.

משפט סופי: ברוכים הבאים , הכנסו בבקשה לאולם.

במשפט השני למילה הראשונה נתננו :

```
positive=['מתכוונת','מוכנה']
```

ולמילה השנייה רק נתנו את המילה עצמה אבל לקחנו את המילה השלישית ב topn = 3

משפט סופי: אני מנסה להאריך את המסמך באותם תנאים.

במשפט השלישי למילה הראשונה נתננו :

```
positive=['אור','טוב']
```

ולמילה השנייה:

```
positive=['פותח','מתחיל']
```

משפט סופי: בוקר חם , אני ממשיך את הישיבה.

במשפט הרביעי למילה הראשונה נתננו :

```
positive=['צהריים']
```

כי עם מילה זו הצלחנו להביא את המודל להחזיר מילה מתאימה יותר, כשהתמשנו במילה "שלום" קבלנו הרבה מילות קשורות לפוליטיקה וזה לא מתאים בשפט זה כי המילה שלום במשפט זה משומשת כמו "היי" או מילה להתחלת דיבור מול קהל.

למילה השנייה:

```
positive=['היקר','הטוב','החבר']
```

ולמילה השלישית רק נתננו את המילה עצמה

משפט סופי: יקרים , הערב התבשרנו שחברינו האמין לא ימשיך איתנו השנה הבאה.

לסיכום : אנו חושבים שהשלמנו את המשישה עם תוצאות טובות מאוד, לא מושלמות אבל טובות מאוד.

ענו על השאלות הבאות:

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.

ישראל: המילים הקרובות כוללות מילים כמו "בחיזוקה", "מזוין", ו"צביונה". הן תואמות חלקית את הציפיות, מכיוון שהן יכולות לשקף דיון פוליטי או ביטחוני סביב ישראל. אבל, המילה "סקוטלנד" מפתיעה משום שמדובר בשתי מדינות שונות.

כנסת: מילים כמו "התמרדות", "בקובלנותיהם", ו"בורסה" עשויות לשקף דיונים פוליטיים וכלכליים בכנסת, אך הקשר אינו ברור לחלוטין מהמילים הללו.

ממשלה: מילים כמו "עיר", "הממשלה", "רשות", ו"יושבי" עשויות לתאר דיונים על ניהול עירוני, רשויות ממשלתיות והחלטות מדיניות. התוצאות כאן הם מונחים רלוונטיים וצפויים להופיע באותו הקונטקסט כמו "ממשלה."

חבר: המילים "שחבר", "וחבר", "לחבר" עשויות להצביע על שימוש נפוץ במונח "חבר" בהקשרים שונים, כולל חברי כנסת. מילים כמו "כהן" ו"נסים" הן שמות פרטיים שמשמשים גם כשמות של חברי כנסת.

שלום: המילים הקשורות כמו "יהודי", "פלסטיני", ו"ישראלי" משקפות את הציפיות עבור דיונים על שלום והסכסוך הישראלי-פלסטיני, מה שמצביע על רלוונטיות גבוהה של המודל לסוגיות אלה.

שולחן: מילים כמו "הונחו", "הונחה", ו"יונחו" משקפות את השימוש במונח "שולחן" בהקשרים של הצעות חוק או תכניות פוליטיות שמונחות לדיון, תואמות את הציפיות לשימוש במונח בדיונים בכנסת.

באופן כללי, חלק מהתוצאות תואמות את הציפיות ומשקפות את ההקשר הפוליטי והמשפטי של הקורפוס, בעוד שאחרות עלולות להראות חוסר עקביות או רלוונטיות פחותה. זה יכול להצביע על כך שהמודל לומד מהקורפוס, אך ייתכן שהוא גם מתקשה לזהות דפוסים מדויקים בשל המגוון הגבוה או ההטיות בקורפוס.

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?

אם ניקח שתי מילים שנחשבות להפכים, אנו מצפים שהמרחק בין שני הווקטורים שלהם יהיה קצר, הצפיה שלנו נובעת מההנחה שמילים הפוכות לעיתים קרובות משמשות באותם הקונטקסטים או בדיונים דומים.

3. מצאו זוג מילים שנחשבות להפכים(antonyms) הקיימות בקורפוס שלנו ובידקו את המרחק ביניהן. האם הציפייה שלכם מסעיף 2 מתקיימת עבורן עם המודל שבניתם?

בדקנו את המרחק בין המילים "לבן" ו"שחור", קבלנו: 0.0789688229560852, נראה שהציפייה שלנו אכן מתקיימת. המרחק הקצר בין שני הווקטורים מרמז שהמודל מזהה את שני המושגים כקשורים אחד לשני ומופיעים בקונטקסטים דומים.

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

בחלק מהמקרים, נראה שהמשפטים הקרובים אכן תואמים לציפיות מבחינת הקשר הסמנטי או הנושאים הדומים שהם מטפלים בהם. למשל, החיפוש עבור "ההצעה המקורית היתה 20 מיליון שקל" שהביא למציאת המשפט "עלות ההצעה היא 4 מיליוני שקלים" מדגים איך שני משפטים העוסקים בנושא של הערכת עלות נתפסים כקרובים במרחב הווקטורי.

מצד שני, ישנם משפטים כמו "אפרת רותם אני ממרכז נגה" והמשפט הכי קרוב "אני מרים ידיים" שלא נראה שיש להם קשר סמנטי ישיר או הקשר מובהק, חוץ מהשימוש בכינוי "אני". זה עשוי להצביע על חוסר בתיאום מסוים בין המשפטים הכי קרובים שהמודל מצא, לבין הקונטקסט או המשמעות המקורית של המשפטים.

הסיבה לתופעה זו נובעת מהאופן שבו המודל Word2Vec עובד. הוא לומד להבדיל ולזהות קשרים בין מילים ומשפטים בהתבסס על ההקשר של השימוש בהם בטקסטים גדולים. זה אומר שהדמיון בין משפטים נקבעים לפי הקונטקסטים שבהם המילים והביטויים מופיעים, ולאו דווקא לפי הגיון או רלוונטיות תוכן סמנטית ברורה.

חלק ג': סיווג

Chunk size = 1:
KNN Stratified Train-Test Split Evaluation:

	precision	recall	f1-score	support
0	0.61	0.65	0.63	3918
1	0.63	0.59	0.61	3917
accuracy			0.62	7835
macro avg	0.62	0.62	0.62	7835
weighted avg	0.62	0.62	0.62	7835

Chunk size = 3:
KNN Stratified Train-Test Split Evaluation:

	precision	recall	f1-score	support
0	0.68	0.70	0.69	1306
1	0.69	0.67	0.68	1306
accuracy			0.68	2612
macro avg	0.68	0.68	0.68	2612
weighted avg	0.68	0.68	0.68	2612

Chunk size = 5:
KNN Stratified Train-Test Split Evaluation:

	precision	recall	f1-score	support
0	0.71	0.71	0.71	784
1	0.71	0.70	0.71	783
accuracy			0.71	1567
macro avg	0.71	0.71	0.71	1567
weighted avg	0.71	0.71	0.71	1567

ענו על השאלות הבאות:

1. האם עבור אותם פרמטרים ותנאים שהשתמשתם בהם בתרגיל 3 (צ׳אנק בגודל 5,אותה כמות שכנים ,שיטת חלוקה וכו)׳ קיבלתם תוצאות טובות יותר או פחות עבור וקטור המאפיינים הנ״ל?

נראה שהשימוש בווקטורי מאפיינים של משפטים (Sentence Embeddings) הביא לתוצאות טובות פחות מאשר השימוש ב-TF-IDF בתרגיל 3. עבור אותם פרמטרים ותנאים, דיוק הסווג הגיע ל-0.71 בממוצע, לעומת דיוק של 0.84 בשימוש ב-TF-IDF.

2.עבור התשובה שעניתם בסעיף 1,הסבירו מדוע לדעתכם זה קרה.

ההבדל בתוצאות יכול להיות מוסבר מכמה סיבות. ראשית, TF-IDF מספק ייצוג טקסטואלי שמדגיש את חשיבותם היחסית של מילים. זה גורם חשב שיכול ליעיל את המודל, במיוחד עבור טקסטים שבהם מילים מסוימות נושאות משקל גדול בהבחנה בין קטגוריות. לעומת זאת, Sentence Embeddings מנסים לתפוס את המשמעות הכוללת של המשפטים, אך עשויים לאבד מידע חשוב על חשיבות יחסית של מילים מסוימות.

3. עבור איזה גודל צ׳אנק קיבלתם תוצאות יותר טובות? האם זה נכון גם לגבי וקטורי המאפיינים שהשתמשתם בהם בתרגיל 3?הסבירו.

התוצאות מראות שעבור גודל צ'אנק גדול יותר,נקבל תוצאות יותר טובות.לכן עבור גודל של 5 קבלנו התוצאות הכי טובות. וזה כן נכון גם לגבי וקטורי המאפיינים שהשתמשנו בהם בתרגיל 3 (TFIDF).

השיפור בדיוק עם הגדלת גודל הצ'אנק מדגים את היתרון של טיפול בקטעי טקסט גדולים יותר, שכן זה מאפשר למודל לזהות טוב יותר את ההקשר הכללי ולהבחין בין הקטגוריות באופן יעיל יותר.

חלק ד׳ : שימוש במודלי שפה גדולים

1.האם קיבלתם משפטים הגיוניים? מבחינת התוכן, קוהרנטיות ומבחינה תחבירית.
המשפטים שהופקו על ידי DictaBERT נראים הגיוניים מבחינה תוכנית ,קוהרנטיות ותחבירית. המודל הצליח להשלים את המילים החסרות באופן שמשמר את משמעות המשפט.

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

כן, יש שיפור משמעותי בתוצאות עבור כל המשפטים בהשוואה לתרגיל בית 2.

3. האם יש משפטים שעבורם המודל עבד פחות טוב? אם כן, הסבירו מה לדעתכם הסיבה לכך. אם לא, האם לדעתכם הוא יעבוד בצורה מושלמת על כל משפט מתוך קורפוס הכנסת? הסבירו.

לא, למרות היכולת המרשימה של DictaBERT לזהות ולחזות מילים בהקשרים שונים, קשה להניח שהוא יעבוד בצורה מושלמת על כל משפט מתוך קורפוס הכנסת כי יש כמה אתגרים שצריך להתגבר עליהם כמו: ניבים ספציפיים, וביטויים תרבותיים או פוליטיים המיוחדים לקורפוס. בנוסף, בקורפוס הכנסת ייתכנו מקרים של מילים בערבית הנכתבות באותיות עבריות כמו: "כלאם פאדי", דבר המצריך הבנה עמוקה יותר של הקונטקסט התרבותי והלשוני.