**דו"ח תרגיל בית 4**

Word Embeddings

**שמות המגישים + ת"ז:**

עוביידה חטיב, 201278066

מאיה עטואן, 314813494

**חלק א'**

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

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

1. הוא מתחיל באות עברי.
2. הוא מסתיים באות עברי או בגרש (').
3. האותיות שבאמצע (חוץ מהראשונה והאחרונה) מורכבות או מאותיות עבריות או מגרש (') או גרשיים (")

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 ב- [',"] ומסתיימת בסימן זהה, והמילה שבאמצע (לא כולל האותיות הראשונה והאחרונה) עונה לשלושת התנאים שלמעלה, גם כן היא נחשבת למילה ורק החלק שבאמצע מתווסף כטוקן לרשימה.

1. 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 הרבה פעמים נוטה לכלול מלים שלא כך קשורות.

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 המשתנה 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 טוב, לעומת וקטור קצר.
* שימוש בווקטור גדול עלול לייצור Overfitting יותר מאשר שימוש בווקטור קטן.
* ווקטור גדול מצריך יותר זמן אימון וחישוב לעומת ווקטור קטן.
* שימוש בווקטורים גדולים מצריך יותר מקום לעומת ווקטור קטן.

1. 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 הקשרים יותר נפוצים מאחרים.

**חלק ב'**

1. 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.
2. הנדרש מבוצע בקוד שלנו ע"י הפונקציה 'sentences\_embed', אשר מקבלת כקלט את המשפטים כרשימה של מלים בלבד (שהתקבלו לאחר ניקוי של הלא-מלים), ואת וקטורי המלים של המודל שחושבו קודם לכן. עבור כל משפט, הפונקציה בונה וקטור 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 הוקטורים שיצרה עבור כל משפט.
3. המשפטים בנחרו כך שיכילו מגוון רב של תכונות. המשפטים לקוחים גם מפרוטוקולי הועדות וגם מפרוטוקולי המליאות, בעלי אורכים שונים, עוסקים בנושאים שונים, משפטים אשר מכילים רק טוקנים שכיחות ואחרים שמכילים טוקנים נדירים, משפטים המכילים עובדות או לא, משפטים בעלי אופי חיובי ומשפטים בעלי אופי שלילי, משפטים נאמרים בגופי זכר, נקבה, ורבים, ומשפטים המכילים 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 ווקטורי ה- Embedding עבור כל אחת, ובאמצעות cosine\_similarity מוצאות עבור כל אחד מהמשפטים שנבחרו את המשפט הכי קרוב מתוך משפטי הקורפוס, ומדפיסה את הפלט בפורמט הנדרש לקובץ ' knesset\_similar\_sentences.txt'.

1. כגדכגד