NLP – HW2

מגישים:

מוחמד חג'אזי – 213011026

מועאד עבד אללטיף - 209418896

**קובץ ה-csv שעליו הסתמכנו הוא זה שאנחנו יצרנו בתרגיל בית 1.

**מימשנו את התוכנית שלנו כך שהיא תרוץ בעזרת ה- cmd דרך הפקודה:

python knesset_language_models.py 'csv_file_path' 'masked_sentences.txt'

כך ש-'csv_file_path' יוחלף בכתובת קובץ ה-csv שלנו(ה- path), וה- 'masked_sentences.txt' גם הוא הכתובת(ה- path) של קובץ ה- masked_sentences.txt. ואז אחרי הרצת התוכנית דרך הפקודה הזו, יוצר שני קבצי פלט(כפי שנדרש) בתקיה שבה נמצא הקובץ knesset_language_models.py.

שלב 1:בניית מודלי שפה

בעת מימוש האינטרפולציה הליניארית בדגם השפה, בחרנו במקדמים $\lambda_1=0.6$, $\lambda_2=0.3$, $\lambda_3=0.1$ על מנת לאזן בין הסבירויות של trigram, bigram ו unigram, עם דגש על החלקת לפלס ליוניגרם. בחירה זו מבטיחה שהדגם יקצה סבירות שאינה אפסית גם כאשר הוא מתמודד עם ביטויים שלא נצפו במידע האימון, תוך שמירה על גמישות ודיוק בחיזוי מבני משפטים.

בחירת המקדמים נותנת חשיבות גדולה להסתברות הunigram בחילוק לפלס, תוך כדי שמירה על משקל ניכר להסתברויות הbigram וtrigram. ההחלטה על חלוקת המשקלים הזו מבוססת על ההנחה שבהינתן טקסט בעברית, יש חשיבות רבה להקשר הרחב (unigram) מאשר להקשר הספציפי הקרובמה שמגביר את הסתברות של הדגם להיות מוצלח בחיזוי מילים במשפט. הבחירה בערכים אלה מטופלת באופן שמאפשר גמישות ורגישות לתדירויות שונות של מילים וביטויים בתוך הקורפוס.

לגבי התמודדות עם משפטים בעלי פחות משלושה טוקנים, הקוד מותאם להתמודד עם זה באופן דינמי. למשפטים עם שני טוקנים משתמשים בהסתברויות bigram, ולמשפט בעל טוקן יחיד משתמשים בהסתברויות unigram. מנגנון זה מבטיח שהדגם ישמר על תפקודו ודיוקו גם למשפטים קצרים יותר.

שלב 2:קולוקציות

מימוש הפונקציה get_k_n_collocations, אנו מחשבים ומדווחים על קולוקציות באורך n עם הכי גבוה Pointwise Mutual Information (PMI) סקורים מתוך k הראשונים. הפונקציה מחשבת את תדירויות האוניגרם וה-n-גרם בקורפוס, ומשתמשת בתדירויות אלו כדי לחשב את PMI לכל n-גרם. PMI מודד את הסיכוי ששני או יותר מילים יופיעו יחד בהשוואה להופעה בנפרד שלהם, תוך שימוש בלוגריתם בסיס 2. ליחשוב הPMI השתמשנו במשואה הזו :

$$
\begin{aligned}
I(x,y) &= \log_2 \frac{P(xy)}{P(x)P(y)} \\
&= \log_2 \frac{P(x|y)}{P(x)} \\
&= \log_2 \frac{P(y|x)}{P(y)}
\end{aligned}
$$

שלב 3 –יישום מודלי השפה

מימשנו פונקציה שקראנו לה find_token כך שהיא מקבלת משפט, ואז בהתחלה עוברים על המשפט ומחפשים את המקומות של הטוקנים החסרים, כלומר מחפשים את [*], שמים את האינדקסים של הכוכביות במערך ואז עבור כל אינדקס בודקים אם הוא קטן מאחד (כלומר המילה הראשונה במשפט חסרה), אם כן אז מחזירים את הטוקן הכי נפוץ במודל הזה, כיוון שאין מילים קודמות כדי להסתמך עליהן. אחרת בודקים אם האינדקס הוא אחד(כלומר המילה השנייה במשפט חסרה), אז משתמשים במילה הראשונה של המשפט ואז קוראים לפונקציית generate_next_token כך שבמקרה כזה היא משתמשת במודל ה- bigram ומחזירה את הטוקן הכי מתאים לבוא אחרי המילה הראשונה. אחרת, אם שני התנאים הקודמים לא התקיימו, זה אומר שהטוקן החסר נמצא בין {2,…,n} ואז גם כאן משתמשים בפונקציית generate_next_token שבמקרה הזה היא מסתכלת על ה- trigram ומחזירה את הטוקן עם הסבירות הכי גבוהה להיות הטוקן החסר.

שלב 4 –שאלות סיכום

1) האם שמתם לב להבדל משמעותי בין שני המודלים שבניתם? האם לרוב קיבלתם את אותן תוצאות בשניהם או תוצאות שונות? הסבירו מדוע לדעתכם זה קרה

בהתבסס על התוצאות שקבלנו בשלב 3, ניתן להסיק שהיו הבדלים בין המודלים שבנינו. על פי התוצאות, נראה שרוב המשפטים שנוצרו נחשבו כיותר סבירים להופיע בקורפוס המליאה, למעט מקרים בודדים שבהם המשפטים נחשבו כיותר סבירים לקורפוס הוועדה. ההבדלים בין המודלים ובין התוצאות שהם יצרו יכולים להיות מוסברים על ידי מספר גורמים:

- אוצר מילים ושפה ייחודיים: כל אחד מהקורפוסים כולל אוצר מילים וביטויים שונים, המשקפים את הסגנון והנושאים הייחודיים להם. מודל שנבנה על בסיס קורפוס מסוים יפיק משפטים שמשקפים את המאפיינים הלשוניים של אותו קורפוס.
- הקשר ותדירות: ההקשרים בהם מופיעות מילים וביטויים בכל אחד מהקורפוסים משפיעים על התדירות ועל הסבירות שמודל יחזה את המשך המשפט. מכיוון שההקשרים שונים, המודלים נוטים לייצר תוצאות שונות בהתאם.
- מבנה המשפטים והתוכן: המשפטים בוועדות ובמליאות עשויים לעסוק בנושאים שונים ולהשתמש במבנים שונים, מה שמשפיע על סגנון השפה והמבנים הסינטקטיים שהמודלים נוטים לייצר.

במקרים שבהם המודלים נתנו תוצאות דומות, זה יכול להיות משום שקיימים ביטויים ומבנים לשוניים שהם נפוצים ורלוונטיים לשני הסוגים של טקסטים. לעומת זאת, ההבדלים בתוצאות מדגישים את החשיבות של התאמת המודל לקורפוס הספציפי עליו הוא מיושם, כדי להגדיל את דיוקו ואפקטיביותו ביצירת טקסט רלוונטי והגיוני (למרות זות ראינו קבלנו כמה משפטים לא הגיוניים – נרחיב בשאלה 3).

בהתבסס על הקולוקציות הנפוצות ביותר(נרחיב בשאלה 2) שנמצאו בקורפוסים של הוועדה והמליאה, ניתן להסיק שוב שקיים הבדל משמעותי בין שני המודלים שבנינו. ההבדלים בקולוקציות מדגימים את השוני בנושאים ובדיונים שמתקיימים בכל אחד מהפורומים האלו.

2) האם הקולוקציות הנפוצות ביותר בכל קורפוס יכולות לספר לנו משהו על התוכן והנושאים בהם הקורפוס עוסק? האם הופתעתם מהתוצאות שהתקבלו או שהן תאמו לציפיות שלכם? הסבירו

הקולוקציות הנפוצות ביותר שנמצאו בכל אחד מהקורפוסים (וועדה ומליאה) אכן מספרות לנו הרבה על התוכן והנושאים שהקורפוסים עוסקים בהם. כל קורפוס מגלה אוצר מילים וביטויים ששכיחים בו ומשקפים את ההקשרים הייחודיים לו .

בקורפוס הוועדה, ניתן לראות שימוש בביטויים המקשרים לנושאים מסוימים כגון "לשגר תנחומים", "מצדדים בחיזוק", ו"שחוקים שמקנים", המצביעים על דיונים פוליטיים וחוקתיים ספציפיים שנערכים בוועדות. הקולוקציות משקפות דיונים טכניים, הצעות חוק, ופעולות פוליטיות שמתבצעות ברמה הוועדתית.

לעומת זאת, בקורפוס המליאה, הקולוקציות כוללות ביטויים כמו "בהתייוונות ובהתבוללות", "המורמונים ביוטה", ו"להתערטל מדתי", המצביעים על דיונים עם רקע תרבותי, דתי וחברתי רחב יותר. השימוש בקולוקציות אלו מצביע על התמודדות עם נושאים גלובליים, תרבותיים ופילוסופיים שנערכים בפלטפורמה הרחבה יותר של המליאה.

ההבדלים בין הקולוקציות בשני הקורפוסים אינם מפתיעים, מכיוון שהם משקפים את הטבע השונה של הדיונים והמטרות של כל פורום. התוצאות תואמות את הציפיות שקורפוס הוועדה יתמקד בנושאים פוליטיים וחוקתיים ספציפיים, בעוד שקורפוס המליאה יכיל דיונים רחבים יותר עם נושאים גלובליים ותרבותיים. התוצאות מספקות תובנות עמוקות לגבי הדינמיקה התרבותית והפוליטית שמשפיעה על השפה והנושאים שנדונים בכל אחד מהפורומים האלו.

3) האם קיבלתם משפטים הגיוניים בחלק 3? פרטו

בהתבסס על התוצאות שהוצגו משלב 3, ניתן לראות שהמודלים שבנינו יצרו משפטים שחלקם היו הגיוניים וחלקם פחות. המודלים הצליחו למלא את החסר במשפטים([*]) עם מילים שהן לעיתים רלוונטיות להקשר, אך לעיתים קרובות הבחירה במילים לא התאימה להקשר הסמנטי או התחבירי של המשפט, מה שהוביל לתוצאות שאינן תמיד הגיוניות.

לדוגמה, במשפט "ההצמדה יהיה שיעור לא הוראת שעה", השימוש במילים "יהיה" ו"שיעור" אינו מתאים סמנטית למשפט המקורי ולכן התוצאה אינה הגיונית. דוגמה נוספת היא "עזבי את את 84, אני לא מכיר את חוק ההסדרים והבנייה", שבה השימוש החוזר במילה "את" לא נכון תחבירית ומוסיף חוסר בהירות למשפט.

לעומת זאת, ישנם משפטים שהמודלים יצרו בהצלחה והם נראים הגיוניים ורלוונטיים להקשר, כמו "אם לא ולא אפשרי את מסתייע, אנחנו נעשה את התיאום במסגרת הוועדה". במקרה זה, המודל הצליח למצוא מילים שמתאימות למבנה ולמשמעות של המשפט המקורי.

התוצאות מראות את האתגרים של יצירת משפטים הגיוניים באמצעות מודלים שפתיים, במיוחד כאשר מדובר במלאי מלים חסרים במשפטים.

4) האם, להערכתם, הייתם מקבלים משפטים טובים יותר או גרועים יותר אם הייתם משתמשים במודלי- BIGRAM ?

ייתכן ששימוש במודלי bi-gram היה מוביל ליצירת משפטים עם איכות נמוכה יותר מאשר עם מודלי trigram, במיוחד במקרים שבהם ההקשר הרחב של המשפט חשוב לניבוי המילה הבאה. עם זאת, במקרים מסוימים, כאשר הטקסט פחות מורכב או כאשר המרחק הסמנטי בין המילים אינו גדול, מודלי bi-gram עשויים להספיק למשימה ולייצר משפטים מספקים.