עיבוד שפות טבעיות

תרגיל בית 2

שם: רנים אבראהים (212920896), אסיל נחאס (212245096)

מבוא:
במהלך התרגיל התבקשנו לבנות וליישם מודלי שפה המבוססים על טריגרמות קורפוס שנלקח מפרוטוקולים של הכנסת. המטרה היא לחשב הסתברויות הופעת משפטים וטוקנים לפי מודל סטטיסטי המבוסס על תדירות הרצפים בקורפוס, וכן לאתר קולוקציות נפוצות ולבצע משימות השלמה וחיזוי טוקנים חסרים במשפט.

הקורפוס מחולק לשני סוגים: משפטים שמקורם בפרוטוקולים מסוג "ועדה" ומשפטים שמקורם בפרוטוקולים מסוג "מליאה". בהתאם לכך, נדרשו שני מודלי שפה נפרדים: אחד לוועדות ואחד למליאות.

שלב 1:

בניית מודלי השפה:
בשלב זה התבקשנו לבנות מודלים המבוססים על טריגרמות. רעיון הטריגרמה הוא שחיזוי הטוקן הבא במשפט מתבסס על שני הטוקנים הקודמים לו.

מחלקת: Trigram_LM
בנינו מחלקה המאגדת את כל המבנים והפונקציות הדרושים למודל השפה: שמרנו במבנה נתונים את ספירות היוני-גראמס (טוקן יחיד), הבי-גראמס (זוג טוקנים רציפים) והטרי-גראמס (שלישיות טוקנים רציפות) עבור כל אחד משני סוגי הפרוטוקולים: ועדות ו- מליאות.

עקבנו אחר מספר הטוקנים הכולל וספירת כל מילה, כדי שנוכל לחשב הסתברויות בהמשך.

הוספנו שני טוקני דמה בתחילת כל משפט, וכן טוקן סיום על מנת לטפל במקרים של תחילת המשפט וסופו לפי הדרישות.

וגם הגדרנו את הלמדות שיהיו המשקל של כל n-gram כך שבמהלך חישוב ההסתברויות של ה- ngrams נרצה לתת יותר משקל ל- unigrams ביחס ל trigrams / bigrams כי ה- unigrams יפחיתו את הרעש ששני ה- ngrams האחרים יוסיפו.

בנינו פונקציה שמחשבת את כמות הטוקנים וגם מצרפת להם טוקני הדמה הפונקציה גם מחשבת את כמות ה- ngrams. הפונקציה נקראת fit_model_to_sentences.

במחלקה שלנו גם קיימות הפונקציות הבאות:

פונקציית: calculate_prob_of_sentence

פונקציה זו מקבלת משפט ומחזירה את לוג ההסתברות שלו לפי המודל.

הוספנו למשפט את הטוקנים הדמה בתחילתו ובסופו.

עברנו על כל טוקן במשפט (החל מהטוקן השלישי) וחישבנו את ההסתברות שלו בהתבסס על שני הטוקנים הקודמים.

ההסתברות חושבה בשיטת נראות מקסימלית עם החלקת לפלס כדי למנוע הסתברויות אפס, ועשינו את החישוב הזה בעזרת הפונקציה compute_smoothed_probability כך שהפונקציה בודקת על פי איזה ngramאנחנו מחשבים את ההסתברות ומציבה בנוסחה שלוקחת את הכמות של ה- ngrams ומוסיפה 1 ומחלקת בכמות הכללית ועוד גודל השפה.

השתמשנו באינטרפולציה ליניארית: שילבנו את הסתברות היוני-גראם, הבי-גראם והטרי-גראם תוך שימוש במשקלים שנבחרו. בתחילה תכננו משקלים שונים, אך על פי ההנחיות ניתן לבחור משקלים ולהסביר בדו"ח. בדוגמה שלנו בחרנו משקלים כך שה- unigram מקבל משקל גדול (0.99), ה-bigram משקל קטן (0.003) וה- trigram משקל בינוני (0.007), במטרה להסתמך יותר על ההקשר הקצר (טוקן אחורה) והסיבה לכך היא כי לפעמים לא רוצים להסתמך על הקונטקסט הארוך ולכן נותנים משקל גבוה ל- unigrams וגם נותנים משקל שהוא קטן ל- unigrams כי לפעמים נרצה לתת משמעות רק לטוקן אחד ולא לקונטקסט רחב כדי להפחית את "הרעש" שה- bigrams יכולים להוסיף

לבסוף סכמנו את לוג ההסתברויות של כל הטוקנים במשפט וקיבלנו לוג הסתברות כוללת למשפט.

פונקציית: generate_next_token

פונקציה זו מקבלת צירוף של טוקנים וחוזה את הטוקן הבא.

בדקנו אם צריך להוסיף טוקני דמה ובחישוב של ההסתברות לטוקן שרוצים לבחור אותו התעלמנו מטוקני הדמה

גם פה חישבנו הסתברות לכל מילה במילון לפי טריגרמה, בי-גראם ויוני-גראם בתוספת החלקת לפלס שמחשבים בעזרת הפונקציה compute_smoothed_probability שהזכרנו קודם.

בחישוב ההסתברות המקדמים שבחרנו הם אותם מקדמים שהגדרנו במחלקה (Lambda_bigram = 0.003, lambda_unigram = 0.99 lambda_trigram = 0.007) כך שנתנו משקל יותר ל- unigrams כדי לקבל תשובה מדויקת יותר.

לאחר חישוב ההסתברויות עבור כל טוקן, בחרנו את הטוקן בעל ההסתברות הגבוהה ביותר והחזרנו אותו יחד עם לוג ההסתברות שלו.

שלב 2:

קולוקציות:

בשלב זה התבקשנו להחזיר את 10 הקולוקציות הנפוצות ביותר באורכי 2,3,4 בכל אחד משני סוגי הקורפוסים (ועדות ומליאות), לפי שני מדדים: (frequency) ו.TF-IDF-

לכן כדי לעשות כך נצטרך לממש פונקציה שמחזירה לנו את k הקולוקציות באורך n הכי נפוצות בקורפוס על ידי מדד מסוים. לכן מימשנו את הפונקציה הבאה:

פונקציית: get_k_n_t_collocations

חילקנו את המשפטים לפי סוג הפרוטוקול: (committee, plenary)

יצרנו n-grams מהמשפטים לדוגמה: (2-grams, 3-grams, 4-grams)

אם המדד הוא ,"frequency" סיננו רק את ה-ngrams שמופיעות לפחות t פעמים, מיינו אותם לפי התדירות בסדר יורד ולקחנו את ה k -המובילים.

אם המדד הוא ,"tfidf" חישבנו TF-IDF עבור כל ngram, סיננו לפי הסף ,t מיינו לפי ערך ,TF-IDF ולקחנו את k הגבוהות ביותר.

את התוצאות הדפסנו לקובץ knesset_collocations.txt בפורמט הנדרש.

תוצאות לדוגמה מתוך הקובץ knesset_collocations.txt:

ניתן לראות בקובץ כי הקולוקציות הנפוצות ביותר (במדד תדירות) כוללות המילים "אני" "זה" וכן מילים מפתח כמו "חבר הכנסת", "היושב - ראש", "אדוני". זה מצביע על כך שבקורפוסים יש הרבה תבניות טקסטואליות שחוזרות (כמו קריאות ביניים, שמות דוברים, סמלי פורמליות).

במדד ה TF-IDF-ניתן לראות הבדלים גדולים: הקולוקציות עם ערכי TF-IDF גבוהים הן עדיין קולוקציות חוזרות, אך המיקום שלהן מבחינת TF-IDF מעיד על כך שהן "ייחודיות" יחסית בתת-קורפוס מסוים.

בפונקציה get_k_n_t_collocations השתמשנו בפונקציות עזר כמו calculate_ngram_frequencies שמחשבת את כמות ה- ngrams במשפטים, ואנחנו נשתמש בערך הזה כדי לחשב כדי לחשב את ערך ה tfidf, ולחישוב ערך ה-tfidf נשתמש גם בפונקצית עזר compute_tfidf שמחשבת הערך על ידי הצבת הארגומנטים המועברים לפונקציה כמו ngram, ngram_frequencies, document_frequencies, num_documents בנוסחת המדד ולאחר החישוב הפונקציה מחזירה את התוצאה.

שלב 3:

יישום מודלי השפה:

פונקציית: mask_tokens_in_sentences

בחרנו אקראית 10 משפטים מתוך קורפוס הוועדות, כל משפט באורך של לפחות 5 טוקנים. מסכנו 10% מהטוקנים בכל משפט (לפחות טוקן אחד) בעזרת "[*]". את המשפטים המקוריים כתבנו לקובץ original_sampled_sents.txt ואת הממוסכים כתבנו לקובץ masked_sampled_sents.txt.

ניבוי הטוקנים החסרים: (generate_results)
לאחר שמסכנו את הטוקנים, השתמשנו במודל המליאה (plenary) לחזות מה הטוקן החסר. למשל, אם המשפט היה "אדוני [*] אתה שאלת , אני אענה*", המודל ניסה לנבא את המילה במקום "[*]"* בקובץ התוצאות sampled_sents_results.txt ניתן לראות שחזה ",", כנראה כי "אדוני , אתה שאלת , אני אענה" הייתה אופציה שהמודל העריך כבעלת הסתברות גבוהה יחסית.

כמו כן, חישבנו את ההסתברות של המשפט המשוחזר בכל אחד משני המודלים (מליאה וועדה) והדפסנו את התוצאות. ניתן לראות שהסתברויות הלוג משתנות, ולעיתים המשפט מתאים יותר למודל המליאה ולעיתים למודל הוועדה.

חישוב:Perplexity
חישבנו את ה Perplexity -עבור הטוקנים הממוסכים בלבד. Perplexity הוא מדד לאיכות המודל: ככל שהוא נמוך יותר, כך המודל מתאים יותר לנתונים.
בקובץ perplexity_result.txt התקבל ערך גדול מאוד (1024.50). משמעות הדבר היא שהמודל, במימוש הנוכחי לא מנבא היטב את הטוקנים הממוסכים. זה צפוי, שכן המודל הוא די בסיסי. ראינו עם הרבה ניסויים גם שככל אנחנו משנים את הלמדות ונותנים משקל יותר גבוה ל unigrams קיבלנו perplexity יותר טובה אבל הגענו למצב שבו הלמדות החלו להשפיע באופן רע על ה perplexity.

הנוסחה שהשתמשנו בה היא: 2^H כך ש: H = 1/N * log(P(W1,…,Wn).

הסבר על הבחירות והמשקלות:
בחרנו משקלים (λ1 = 0.007, λ2 = 0.003, λ3 = 0.99) המעדיפים באופן כבד את unigrams, בתקווה שהקונטקסט הקצר יותר יסייע לסיבוכיות קטנה.

גם ראינו שכאשר השתמשנו בלמדות שנותנות העדפה פחות ל unigrams קיבלנו ערך perplexity גדול יותר ממה שקיבלנו עם הבחירה של lambda_unigrams = 0.99 וגם כאשר נסינו לתת משקל יותר גבוה, למשל: lambda_unigrams = 0.999 מצאנו שגם ערך ה perplexity עלה, לכן מפני המשקלים שבחרנו נתנו התוצאה הכי טובה דרך trial and error, החלטנו לקחת אותם להיות הבחירה שלנו.

שלב 4:

שאלה 1:

הקולקציות על פני התדירות יש להם משמעות חלקית ויכולות כמעט לספר לנו משהו על תוכן הקרפוס, אם אנחנו מסתכלים על 2 מילים ו 3 מילים בשני הקורפסים:

א. קורפוס הוועדות: הופיעו לנו מילים כמו "אני," , "אני רוצה","אנחנו," אלו קולקציות שהם כללים אשר לא נוכל להבין שום דבר על תוכן הקמפוס חוץ משהוא מהווה דו שיח.

ב. קורפוס המילאה: הופיעו לנו מילים כמו: "חבר כנסת" , "אדוני היושב" , "אינו נוכח" קולקציות אלה ניתן להבין שהקורפוס מדבר על דיונים בכנסת.

אם מסתכלים על four grams collocations אנחנו נראה שעכשיו בקורפוס הוועדות אנחנו יכולים קצת להבין שמדובר לדיונים עם יושב ראש מצירופים כמו: " אדוני היושב – ראש". אם אנחנו מסתכלים על צירופים מקורפוס המילאה עם 4 מילים נוכל להבין יותר מצירופים בשני מילים על מה מדבר הקרפוס לדוגמה: " רבותי חברי הכנסת ," אדוני היושב – ראש".

האם הופתענו מהתוצאות? לאמת שלא, כי ראינו שהמילים האלו מופיעות הרבה מאוד בקורפוס ולכן היה סביר להניח שהן יהיו הקולוקציות שייבחרו.

שאלה 2:

לפי הגדרת tf-idf היא משקפת את מרכזיותה של הקולקציה במסמך, ביחס למסמכים אחרים. אם נסתכל על הקולקציות שמתקבלות משני מילים נבין חלקית על מה כמה מסמכים דיברו ביחס למסמכים אחרים כמו: " הטלוויזיה החינוכית" , " התיישבות היהודים" אם נסתכל על צירופים משלוש וארבע מילים נראה מה שראינו אך בצורה יותר ברורה וחדה למשל: " רשות הדיבור לחבר הכנסת", " פרק ב לחוק התיאומים" ו " במקרקעין המצויים מחוץ לישראל" אך הופתענו שלמרות היחודיות של כמה צירופים הופיעו גם צירופים כללים בעיקר בקורפוס המליאות למשל: ״, חברי כנסת,״.

שאלה 3:

כן ראינו הבדלים בין שני המדדים הסיבה מכך מגיעה מההגדרה של כל אחת מהם: בקרפוס הראשון היא נפיצות הביטוי בכל הקורפוס והשני מרכזיותה של הקולקציה במסמך, ביחס

למסמכים אחרים. כלומר, אנחנו לא נמצא ביטוים שמצאנו בתדיריות כמו "אני רוצה" או "אני," כי במדד השני בכל דיון יכולים להופיע כל אחת מהביטוים כי אלה חלקי דיבור עיקריים ולכן הן יהיו נפוצים בכל הקבצים ולכן הם לא הערך שלהם על פי מדד השני יהיה נמוך. אך לפעמים ראינו דמיון בעיקר בקורפוס של המליאה וזה היגוני משום שבמילאה עוסקים בנושאים שלא יכולים להיות מגונים כמו בועדות אשר מתמקדים בנושאים רבים.

שאלה 4:

ה- threshold כאשר אנחנו מסתכלים על מדד התדירות הוא נותן לנו ה- ngrams הכי נפוצות אם אנחנו מגדילים את הערך שלו ולהפך גם אם נפחית את ה- threshold נקבל שהוא אפשר ל- ngrams שהן פחות נפוצות להיבחר.

כאשר אנחנו מדברים על מדד ה- tfidf אז אם נגדיל את ערך ה- threshold אז נקבל ngrams שהן ייחודיות אך מופיעות הרבה בקורפוס, ואם נפחית את הערך שלו אז גם נקבל ngrams שהן ייחודיות אבל שמופיעות פחות בקורפוס.

שאלה 5:

התוצאות שהתקבלו היו מגוונות מבחינת ההיגיון שלהן.

שאלה 6:

חלק מהמשפטים שהושלמו היו הגיוניים מאוד ואחרים היו פחות הגיוניים, ציפינו תוצאות אלה משום שהחיזוי תלוי בטוקנים שנבחרו ומראה גם שהוא תלוי בנפיצות גבוהה של טוקנים כמו הפסיק, וזה יכול להוביל לתוצאות שאינם תאומות לתשובה הנכונה. לדוגמה: במשפט "אדוני , אתה שאלת , אני אענה" החיזוי היה נכון, אך במשפט " אנחנו מאוד רגישים לנושא הזה ואנחנו לא נאפשר שילד שירצה להיות בפעילות שלנו הוא לא יהיה בפעילות שלנו – זה בכל תנועות הנוער . החיזוי היה שגוי. המודל העדיף מילים עם נפיצות גבוהה, למרות שהן אינן נכונות דגוגית במשפט.

שאלה 7:

נוסחת ה-perplexity מודדת עד כמה המודל מצליח לחזות מילים חסרות. אם הערך נמוך משמעות זה חיזוי טוב (ההיפך נכון). במילים אחרות אם החיזוי נכון אז ההסתברות גבוהה והלוג קרוב ל-0, מה שמוביל ל-perplexity נמוך. ואם החיזוי שגוי אז מתקיים ההיפך.

שאלה 8:

אחרי כמה נסיונות למדידת ה perplexity- התוצאה לא הייתה טובה במיוחד הסיבה מכך שהמודל מנחש באופן סטטיסטי הוא לא למד דגדוג של השפה וגם לא היה קורפוס מספיק כדי לשקף שימוש מגוון ועשיר בשפה ולכן, ערכי ה-perplexity עלו.