דו"ח תרגיל 2 – עיבוד שפות טבעיות

מגיש: שגיא גוילי, ת.ז. 203638804

בתרגיל זה התבקשתי להכין תוכנית אשר מקבלת כקלט תיקיית קבצי טקסט, אשר מכילים ערכים מויקיפדיה שיופיעו בשפה הספרדית ו/או אנגלית ו/או אנגלית פשוטה, מאגדת כל ערך תחת קורפוס השפה המתאים לערך ובסופו של דבר מייצרת 3 קורפוסים- אחד לכל שפה.

כל קורפוס אני מחלק לשלוש טבלאות (יופיעו כמילונים"[Dictionary] בתוכנית), טבלת טוקנים, צמדים ושלשות, כאשר כל טבלה תכיל את מספר ההופעות של כל איבר בה ואת הסתברותו.

בנוסף, כפי שנדרש בהוראות, אני שומר את תוכן הקורפוס עצמו עם סימן התחלה לכל תחילת משפט (<s>) וסימן סיום לכל סוף משפט (</s>) כאשר ההנחה היא שכל שורה בקובץ טקסט של ערך מהווה משפט.

שלבי יצירת הטבלאות:

Unigrams:

כל "טבלה" היא בעצם רשימה המכילה שני מילונים- האחד סופר הופעות של כל טוקן והשני הסתברות של כל טוקן. שיטת הספירה היא פשוטה למדי- המפתח במילון הספירה הוא הטוקן עצמו והערך של המפתח הוא מספר הפעמים שהוא הופיע. באשר לשיטת החישוב ההסתברות היא כפי שהוסבר בהוראות- מספר ההופעות של כל טוקן חלקי אורך הקורפוס (כלומר מספר הטוקנים הכולל בקורפוס)

Bigrams:

בדומה ל-Unigrams, גם פה אני שומר את "טבלת" ה-Bigrams בשני מילונים, כאשר סופר ההופעות עובד בצורה דומה, אך בשונה משיטת חישוב ההסתברות הקודמת, במודל זה אני מבצע החלקה לפלסיאנית (Laplace Smoothing) לכל צמד, כלומר אני מחלק את $C(w_{n-1}w_n) + 1$ (כלומר את מספר המופעים של הצמד בקורפוס ועוד 1) במספר המופעים של $C(w_{n-1}) + V$ (כלומר במספר המופעים של הטוקן הראשון בצמד ועוד מספר **סוגי** הטוקנים בקורפוס) ולאחר מכן, לפי כלל השרשרת, אני כופל את המנה בהסתברות של הטוקן הראשון שחישבתי קודם לכן במודל ה-Unigrams.

Trigrams:

בדומה לשני המודלים הקודמים גם במודל זה אני שומר את "טבלת" ה-Trigrams בשני מילונים, כאשר סופר ההופעות עובד בצורה דומה (עובר שלשה שלשה בקורפוס ומוסיף 1 למפתח המתאים לה במילון ספירת ההופעות), אך בשיטת חישוב ההסתברות במודל זה נעשה Backoff על ידי שימוש ב-Simple Linear Interpolation שכפי שנלמד בהרצאה, כאשר בעצם אני מבצע אינטרפולציה בין ההסתברויות של הטוקן האחרון בשלשה, הצמד האחרון בשלשה והשלשה עצמה. ראשית אני מחשב את ההסתברות של השלשה עצמה- $P(w_n|w_{n-2}w_{n-1}) = \frac{C(w_{n-2}w_{n-1}w_n)}{C(w_{n-2}w_{n-1})}$, לאחר מכן אני בוחר את הדלטות (המקדמים) בהם אני כופל כל הסתברות לפני החיבור ביניהן:

$\delta_1 = 0.6, \delta_2 = 0.2, \delta_3 = 0.2$; החלטתי את מירב המשקל במקדמים לתת להסתברות השלשה עצמה, מכיוון שאם שלשה מופיעה ברצף בקורפוס בהסתברות נמוכה אזי אני רוצה לשמור על הסתברותה הנמוכה (ההסתברות היא שבר ברוב המקרים ולכן הכפלה בשבר גדול בעצם תקטין אותה), להסתברות של הצמד האחרון חשבתי לתת הסתברות גדולה יותר מההסתברות של הטוקן הראשון מתוך מחשבה שכמו בשלשה- אם הצמד מופיע בהסתברות נמוכה בקורפוס ארצה לשמור על הסתברות זו נמוכה אבל לאור ההנחייה לבצע החלקה ל-Bigrams ההסתברות גם ככה ירדה יותר מההסתברות המקורית ולכן לבסוף החלטתי לתת להסתברות הטוקן האחרון ולהסתברות הצמד האחרון את אותה הדלטא.

שלב שיוך המשפטים לקורפוס המתאים:

בחלק זה של התוכנית הגרלתי בצורה רנדומלית משפטים באורכים המבוקשים בסעיף א' של כל מודל בהוראות התרגיל. בכדי לשייך כל משפט לקורפוס המתאים לו השתמשתי בכלל השרשרת. הבעיה המרכזית שנתקלתי בה הייתה מילים OOV, בכל מודל טיפלתי בבעיה זו לפי צורת העבודה שהתבקשה בתרגיל.

Unigrams:

עבור מודל זה הגרלתי משפטים רנדומליים באורכים 1, 2 , 3, 5 – במקרים של מילים OOV נתתי למילה שלא מופיעה בקורפוס (כלומר OOV) הסתברות 0 שכן הוראות התרגיל לא דרשו טיפול מיוחד לעניין. דוגמאות לשיוך משפטים-

משפטים שאני בחרתי:

אורך 1- " romance " נבחר מתוך קורפוס האנגלית הפשוטה ואכן שויך כמצופה לקורפוס המתאים של האנגלית הפשוטה.
אורך 2- " football field " - נבחר מתוך קורפוס האנגלית ואכן שויך כמצופה לקורפוס המתאים של האנגלית.
אורך 3- " cup of tea " – נבחר מתוך קורפוס האנגלית ואכן שויך כמצופה לקורפוס המתאים של האנגלית.
אורך 5- " it was made in china " – בחרתי צירוף זה מהראש וציפיתי שיקושר לקורפוס האנגלית- הדבר אכן קרה.

משפטים שהוגרלו באקראי:

אורך 1- " based "- הייתי מצפה שישויך לקורפוס האנגלית ואכן הדבר קרה.

אורך 2- " Wikipedia no " - על פניו צירוף זה מכיל שתי מילים שהייתי מצפה לשייך לקורפוס האנגלית או האנגלית הפשוטה, אך אם נחשוב על משמעותו נבין שלהגיד "Wikipedia no" לרוב לא מקובל בדקדוק האנגלי שכן היה מצופה להגיד is not, אולי will not וכדומה.. לכן לאחר שחקרתי את העניין (שכן איני דובר ספרדית) ההקשר פה היה מתאים יותר לקורפוס הספרדית שכן no זו מילת שלילה גם בספרדית ודקדוקית יכולה להופיע בשפה זו- להפתעתי צירוף זה אכן שויך לקורפוס הספרדית למרות שהוא לא מסתכל מילה אחורה אלא רק על המופע הבודד של המילה ולכן ההקשר של הצירוף יחדיו לא היה אמור להוות פקטור, אך ככל הנראה השימוש ב-no בספרדית הוא רב יותר מאשר באנגלית ולכן המילה אכן נבחרה לקורפוס זה.

אורך 3- " tropical cyclone worldwide " – הייתי מצפה שישויך לקורפוס האנגלית או האנגלית הפשוטה, ואכן שויך לקורפוס האנגלית.

אורך 5- " un cráneo más pequeño ; " – מהסתכלות ראשונה ניתן לראות פה מילים בספרדית לכן הייתי מצפה שישויך לקורפוס הספרדית ואכן הדבר קרה.

לסיכום, ניתן לראות שהסתברות 0 למילים OOV אכן פתרה את הבעיה במודל זה והשיוך עבד בצורה מדויקת ויעילה.

Bigrams:

עבור מודל זה הגרלתי משפטים רנדומליים באורכים 2, 3, 4, 5 – במקרים של מילים OOV החלקתי את הטוקן הראשון על ידי הנוסחא $\frac{C(w_i)+1}{N+V}$ (כאשר N הוא מספר הטוקנים בקורפוס ו-V מספר סוגי הטוקנים) ואת הצמד עצמו על ידי שימוש בנוסחא $\frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$ כאשר V הוא מספר סוגי הצמדים בקורפוס, עשיתי זאת מתוך מחשבה שיהיו לי הרבה יותר סוגי צמדים בקורפוס מאשר סוגי טוקנים וכאשר המכנה גדול יותר ההסתברות קטנה יותר והרי אם צמד לא קיים בקורפוס לא ארצה לשייך אותו לקורפוס. לבסוף הכפלתי את שתי המנות לפי כלל השרשרת.

משפטים שאני בחרתי:

אורך 2- " no fear " – בחרתי צירוף זה מהראש בציפייה שהוא ישויך לקורפוס האנגלית או האנגלית הפשוטה שכן מונח זה מאוד מקובל בשפות הללו וכן הוא גם מתאים לכללי הדקדוק בשפה, להפתעתי צירוף זה שויך לקורפוס הספרדית, בדומה ל-"Wikipedia no" ממוקדם אני מניח שמספר ההופעות של המילה no היא זו שהכריעה את הכף שכן צירוף זה הוא ככל הנראה OOV בכל אחד מן הקורפוסים והעיקר המשקל ניתן לטוקן no והסתברותו.

אורך 3- " a rotten apple " - בחרתי צירוף זה מהראש מתוך צפייה במילים בקורפוס הספרדית- יש שימוש רב במילה a ורציתי לראות האם הקורפוס אכן התאמן ולמד לשייך משפט מסוג זה מתוך הקשר של הצמד a rotten ו-rotten apple , לשמחתי הדבר אכן קרה כמצופה.

אורך 4- " del loves eating chocolate " – חשבתי על צירוף זה מהראש בציפייה שישויך לקורפוס האנגלית או האנגלית הפשוטה, שכן למרות ש-del היא מילת קישור נפוצה בספרדית ו-chocolate היא מילה בספרדית גם כן, עדיין דקדוקית הצירוף פחות מתאים לספרדית ויותר לאנגלית ולכן שמחתי לגלות שמודל ה-bigram שייך צירוף זה לאנגלית הפשוטה.

אורך 5- " chipped a bone in his shoulder " – בחרתי צירוף זה מקורפוס האנגלית הפשוטה, יש כאן מונחים שיכולים לדעתי להתאים גם לאנגלית אבל שמחתי לגלות שה-bigram שייך צירוף זה לקורפוס האנגלית הפשוטה כמצופה.

משפטים שהוגרלו באקראי:

אורך 2- " , 1951 "- במקרה זה קשה מאוד להכריע לאן צירוף זה יכול להתאים, בכל שפה שנה נרשמת בצורה הזאת, כך שככל הנראה בגלל שהשנה הספציפית הזו הופיעה יותר בקורפוס הספרדית אזי לקורפוס זה שויך הצירוף.

אורך 3- " stellar disk is " - צירוף זה הייתי נוטה לקשר לשפה האנגלית הפשוטה, אני אישית לא הכרתי את המשמעות של stellar, ולכן הנחתי שהוא מגיע מהאנגלית פחות מוכרת. צירוף זה אכן שויך לקורפוס האנגלית הפשוטה.

אורך 4- " as a reference view " – הייתי מצפה שישוייך לקורפוס האנגלית או האנגלית הפשוטה, ואכן שויך לקורפוס האנגלית.

אורך 5- " con Death of a Ladies " – בהסתכלות ראשונה סביר להניח שצירוף זה יקושר לשפה האנגלית או האנגלית הפשוטה, רוב המילים הן באנגלית פרט למילה הראשונה וזה נשמע צירוף די הגיוני אך ה-con בהתחלה, שהוא מהשפה הספרדית, נצפה ככל הנראה כמגיע לפני המילה Death בקורפוס הספרדי והמילה Death פחות הופיעה במיוחד בצמדים של הצירוף הנ"ל לכן צירוף זה שויך לקורפוס השפה הספרדית.

לסיכום, ניתן לראות שההחלקה פתרה לנו את בעיית חוסר המילים בקורפוס ונתנה מענה יפה שכן רוב הצירופים מתמיינים בצורה טובה ובן אדם ממוצע ככל הנראה היה משייך בצורה דומה.

Trigrams:

עבור מודל זה הגרלתי משפטים רנדומליים באורכים 3, 4, 5, 7 – במקרים של מילים OOV החלקתי (בדומה להחלקה לפלסיאנית) את הטוקן האחרון, את הצמד האחרון ואת השלשה עצמה (עם הנוחסאות שפורטו במודל הקודם, כאשר עבור השלשה השתמשתי בנוסחא $\frac{C(w_{n-2}w_{n-1}w_n)+1}{C(w_{n-2}w_{n-1})+V}$ כאשר כמו בביגרמים גם כאן V הוא מספר סוגי השלשות בקורפוס מאותה הסיבה שפורטה בביגרמים) על מנת שאוכל להשתמש בהם לביצוע של Simple Linear Interpolation עם המקדמים שפורטו בשלב יצירת הטבלאות.

משפטים שאני בחרתי:

אורך 3- " feel no fear " – בחרתי צירוף זה כדי לבדוק את הישנות המצב שתואר במודל ה-Bigram עבור הצירוף "no fear" , שכאמור סיווג צירוף זה לשפה הספרדית, הוספתי לו מילה שהשאירה לנו צירוף הגיוני בשפה האנגלית ופחות בספרדית, הציפייה שלי הייתה שבשלב זה כאשר הצירוף נראה מספיק "אנגלי" והמודל קצת חכם יותר ומסתמך על אימון נרחב יותר, הוא אכן ישויך לקורפוס האנגלית או האנגלית הפשוטה ולשמחתי הוא אכן שויך לקורפוס השפה האנגלית הפשוטה.

אורך 4- " los angeles lakers team " - בחרתי צירוף זה מהראש בניסיון לבחון את יכולות המודל כאשר שתי המילים הראשונות תואמות גם לשפה הספרדית, בעיקר המילה los, וגם לשפה האנגלית מה שיכל לאתגר את המודל – המחשבה שלי הייתה שמודל זה מתאים לשפה האנגלית או האנגלית הפשוטה, "קבוצת הלוס אנג'לס לייקרס"- האדם הממוצע שהיה קורא משפט זה ככל הנראה היה חושב שסביר להניח שזה אכן באנגלית, לשמחתי המודל אכן שייך צירוף זה לשפה האנגלית הפשוטה.

אורך 5- " Death of a Ladies is " – חשבתי על צירוף זה מהראש בהמשך לצירוף שנבחר באקראי במודל ה-Bigrams, בכדי לאתגר את המודל בכך שהמשפט בעצם יכיל מילים באנגלית בלבד וגם דקדוקית הוא יהיה נכון בשפה האנגלית מה שהיה גורם לאדם הממוצע לשייך צירוף זה לשפה האנגלית ללא היסוס רב, אך מכיוון שהקורפוס בשפה הספרדית אומן על הרצף הנ"ל (ללא is) וככל הנראה לא היה מופע של is אחריו- הוא עדיין סיווג צירוף זה לשפה הספרדית מה שמבחינתנו לא נכון אמנם אבל מהאימון שלו זה השיוך המתאים.

אורך 7- " Montreal , Quebec , Israel , Canada " – בחרתי צירוף זה מקורפוס השפה הספרדית, עם שינוי קל - הוספתי Israel, רוב האנשים שהיינו מבקשים מהם לשייך צירוף זה היו משייכים אותו קרוב לוודאי לקורפוס האנגלית או האנגלית הפשוטה- כל המילים באנגלית, נראה כמו פירוט ארצות, אך מכיוון ש-4 טוקנים הראשונים בצירוף מופיעים בקורפוס הספרדית אזי הם הטו את הכף לטובת קורפוס הספרדית- כך שמבחינת הקורפוס ההתאמה זו היא המתאימה ביותר כי Israel לא נראה באף קורפוס ככל הנראה מגיע אחרי Quebec או לפני Canada (עם הפסיקים כמובן)

משפטים שהוגרלו באקראי:

אורך 3- " or early 19th " - בהסתכלות ראשונה הייתי משייך צירוף זה לקורפוס האנגלית או האנגלית הפשוטה ואכן המודל שייך לקורפוס האנגלית.

אורך 4- " == History == The " - צירוף זה אמנם מכיל "==" פעמיים מה שיכול קצת להטיל ספק שכן אלה סימנים הקשורים לכל השפות, אבל המילים מכריעות שהשיוך אמור להיות לשפה האנגלית ואכן המודל שייך אותו לקורפוס השפה האנגלית.

אורך 5- " Sin embargo , a veces " – בצירוף זה ניתן להבחין שקיימות מילים רבות באנגלית, בהסתכלות ראשונית הייתי משייך צירוף זה לקורפוס השפה האנגלית, אך בהסתכלות מעמיקה יותר ניתן להבחין שמבחינת ההקשר קשה להבין את משמעות המשפט באנגלית, בנוסף לכך שהטוקן

האחרון הוא בספרדית והמילים האחרות הן גם כן חלק מהשפה הספרדית הייתי אולי נוטה בסופו של דבר לשייך משפט זה לקורפוס הספרדית ואכן המודל שייך את הצירוף לקורפוס הספרדית.

אורך 7- " Tagliaferro también introdujo el automóvil de Donald " – מכיל ברובו מילים בשפה הספרדית לכן הייתי משייך אותו לקורפוס השפה הספרדית ואכן המודל שייך אותו לקורפוס השפה הספרדית.

לסיכום, ניתן לראות שההחלקה והאינטרפולציה פתרה לנו את בעיית חוסר השכיחות של צירופי המילים הרנדומיים ובמקרים רבים (שלא תיכננתי במדויק כדי לאתגר את המודל) המודל שייך את הצירופים לקורפוסים שכל אדם ממוצע היה משייך גם כן.

שלב יצירת המשפטים האקראיים:

בשלב זה נדרשתי להשתמש קורפוסים המאומנים שהכנתי לכל שפה על מנת לייצר משפטים רנדומליים על ידי שימוש בכל אחד מן המודלים. את אורך כל משפט הגרלתי על ידי שימוש בהתפלגות אורכי המשפטים בכל קורפוס (כלומר במונה ספציפי לכל אורך משפט שהופיע בקורפוס) ואכן קיבלתי משפטים רנדומליים באורכים שונים.

אפרט כעת איך התבצעה יצירת המשפטים בכל מודל:

Unigrams:

מודל זה מבחינתי היה הפשוט ביותר, ראשית הגרלתי אורך כמתואר להעיל, לאחר מכן השתמשתי בהסתברויות של כל טוקן במודל (הערכים של ה"טבלה" כלומר המילון) כדי להגריל באקראי טוקן (המפתחות של ה"טבלה" כלומר המילון) בכל איטרציה עד אשר קיבלתי משפט באורך המוגרל, חשוב לציין שסיננתי תחיליות וסיומות (<s> ו-</s>)
ביצעתי את הפעולה 3 פעמים, כל פעם עבור קורפוס שפה אחר.
ניתן להבחין שהמשפטים שיצאו במודל זה יצאו ברוב המקרים מאוד לא מובנים, קריאים או הגיוניים שכן אין פה בכלל קשר למילים שבאו לפני- כל מילה נבחרת על סמך מספר המופעים שלה בקורפוס בלבד ותו לא. דבר נוסף שמתקבל כתוצאה מכך הוא הרבה מאוד משפטים לא מסודרים מבחינת סימני הפיסוק: פסיק אחרי פסיק, סוגר ללא פותח, פותח ללא סוגר, נקודה שאחריה עוד רווח ונקודה וכדומה

Bigrams:

במודל זה הכנסתי אלמנט קצת יותר מתוחכם מצורת בחירת המילים שהייתה ב-Unigrams, ראשית הגרלתי את האורך כמובן, לאחר מכן הכנתי רשימה של כלל הצמדים שמכילים את התחילית שהוספתי לכל תחילת משפט (<s>), כלומר רשימה המכילה צמדים עם תחילית ואחריה מילה אחת שהיא אכן מהתוכן המקורי של הערך. כל צמד שכזה שמרתי כמפתח במילון שיצרתי כאשר הערך של כל מפתח הוא ההסתברות של אותו הצמד במילון הצמדים המקורי של המודל. כעת, הגרלתי מהרשימה הזו בלבד את הצמד הראשון שיופיע במשפט האקראי (כאשר את התחילית עצמה, <s>, לא הכנסתי למשפט אלא המילה הראשונה הייתה בעצם הטוקן השני בצמד שנבחר באקראי).

בצורה הזו וידאתי שכל משפט שאני יוצר הוא אכן התחלה של משפט אמיתי שהיה בערכי האימון עליהם אימנתי את הקורפוס והמשמעות של המשפט תהיה הרבה יותר הגיונית ותקינה.

כעת בכל איטרציה יצרתי רשימה (שזה בעצם מילון מבחינת המבנה נתונים בפייתון) שמכילה את כל הצמדים האפשריים הבאים בקורפוס שמכילים מילה אחרי המילה שכרגע הוספתי למשפט, לדוגמא אם הוספתי את המילה moshe אזי הרשימה שנוצרת תכיל את כל הצמדים המכילים את moshe כטוקן הראשון בהם בקורפוס. כל צמד יהווה מפתח והערך של כל מפתח יהיה ההסתברות של הצמד הזה ברשימת ההסתברויות המקורית של המודל. ברשימה שנוצרה אני משתמש בכדי להגריל את המילה הבאה שתיכנס למשפט (כמו בצמד שהיה עם התחילית גם כאן אני לא מכניס שוב את moshe, שבדוגמא, אלא רק את הטוקן השני בצמד שנבחר באיטרציה)

אני ממשיך הפעולה זו עד אשר אני מגיע לאורך שהגרלתי בהתחלה או עד אשר אני מגיע לסיומת (</s>) מתוך מחשבה שאם הגעתי לסוף משפט תקין אני מעדיף להשאיר רק אותו גם אם האורך המוגרל היה גדול יותר מאורך המשפט שבסופו של דבר קיבלתי. (גם את הסיומת אני לא מכניס למשפט, כלומר את ה-</s>)

במהלך הריצות הראשונות של יצירת המשפטים במודל זה נתקלתי בשגיאת זמן ריצה שהפילה את התוכנית, אחרי בדיקה של הבעיה שמתי לב שלפעמים רשימת האופציות הבאות ממנה אני מגריל את המילה הבאה נותרת ריקה- ככל הנראה כי לא נמצא צמד למילה שהכנסתי למשפט באיטרציה האחרונה. את הבעיה פתרתי על ידי כך שהתחלתי לבחור מחדש מילים מרשימת הצמדים עם התחיליות (כלומר רשימת הצמדים שהטוקן הראשון שלהם היה <s>) במחשבה שאם אין אפשרות למצוא המשך למילה הנוכחית אז לפחות אתחיל משפט חדש שבסיכוי לא רע יהיה גם הגיוני.

ניתן להבחין שהמשפטים שנוצרו היו יחסית נכונים מבחינת התחביר; אות גדולה בתחילת משפט, נקודה בסוף משפט, קשרים במקומות הנכונים. לצד זה סימני הפיסוק שתוחמים לנו טקסט, סוגריים רגילים ומרובעים למשל, עדיין הופיעו לפעמים רק עם סוגר בלי פותח או להפך.

בנוסף המשפטים עדיין לא היו לגמרי הגיוניים וקשורים, בעיקר אחרי נקודה הגיע משפט חדש שלאו דווקא קשור לקודמו בצורה כלשהיא, אך גם באמצע משפט כשהגיעו הקשרים and, or, with וגם a הרבה פעמים המשפט עבר לנושא אחר בכלל שלא קשור למה שהיה לפני- אני חושב שהדבר נבע בעיקר מזה שמודל זה מסתכל רק מילה אחת אחורה, מה שהגדיל את כמות האופציות לבחירה בכל איטרציה כשהקורפוס הוא גדול מאוד וזה גרר שיש למשל הרבה אופציות למילים אחרי and ורובן לא יהיו קשורות למה שבא לפניו. במקרים רבים הדבר נבע גם מצורת הפתרון לבעיית חוסר הצמד למילה האחרונה שציינתי מקודם, התחלת משפט חדש לעיתים באמצע השורה גרמה לי לאבד את ההקשר שהיה למשפט עד כה.

<u>Trigrams</u>:

במודל זה עבדתי בצורה דומה למה שפורט במודל ה-Bigrams, הגרלתי אורך, ולאחר מכן הכנתי את רשימת/מילון התחיליות, שהפעם מכיל שלשות של תחילית ושתי מילים נוספות אחריה, כאשר באיטרציה הראשונה הכנסתי את שתי המילים האחרונות בשלשה שנבחרה מתוך רשימה זו, לאחר מכן הכנסתי את רשימת השלשות האפשריות הבאות בקורפוס שמכילות את הטוקן האמצעי בשלשה הנוכחית כטוקן הראשון שלהן ואת הטוקן האחרון בשלשה הנוכחית כטוקן השני שלהן, בדרך זו השגתי כל פעם עוד מילה אחת שמתאימה בהקשר לשלשה הנוכחית והמשפט שהורכב נהיה הרבה יותר הגיוני. לאחר שנבחרה השלשה, אני עושה shift שמאלה בשלשה הנוכחית ומוסיף אליה את הטוקן הראשון בשלשה החדשה שנבחרה ובאיטרציה הבאה אני מכניס את הטוקן הראשון בשלשה החדשה שנוצרה. דוגמא להמחשה:

נניח שהשלשה הנוכחית היא I love LA, באיטרציה הבאה אני אכניס את הטוקן "I" למשפט ולאחר מכן אחפש את כל השלשות מהצורה love LA X כאשר X הוא טוקן כלשהו. נניח השלשה שנבחרה היא love LA so אזי זו תהיה השלשה הבאה שלי כאשר באיטרציה הבאה יכנס הטוקן "love" למשפט וכן הלאה.

המשכתי לרוץ על האלגוריתם הנ"ל עד אשר הגעתי לאורך הרנדומלי שהוגרל בתחילה או עד אשר הגעתי ל-</s> מאותן הסיבות שפירטתי ב-Bigrams. כמובן שגם פה סיננתי את <s> ואת </s>.

בדומה ל-Bigrams גם פה קרו לי מקרים בהם הרשימה של האפשרויות הבאות הייתה ריקה- דרך הטיפול בבעיה הייתה דומה לדרך הטיפול ב-Bigrams.

ניתן להבחין שהמשפטים האקראיים שנוצרו במודל זה יצאו הרבה יותר הגיוניים- משפטים באורכים קצרים עם משמעות הגיונית עם סבירות לא רעה שבן אדם ממוצע יכול לכתוב, בעיקר בהשוואה למודל הקודם בו רוב ההיגיון במשפט נגמר אחרי 4-5 מילים. בנוסף המשפטים יצאו הרבה יותר נכונים תחבירית- פסיקים ונקודות מופיעים במקומות הרבה יותר הגיוניים, הרבה פחות סוגריים עם פותח ללא סוגר וסוגר ללא פותח. לדעתי החיפוש של המילה הבאה על ידי הסתמכות בשני מילים

אחורה עשתה פה את השינוי המירבי, למרות שלא הכל מושלם ועדיין ישנם מקרים לא מעטים של טעויות תחביריות וחוסר היגיון לוגי.

שאלות לדו"ח:

שאלה: השוו את המשפטים שקיבלתם בכל אחד משלושת המודלים שהתבסו על כל שפה בנפרד. האם יש הבדלים באיכות השפה? האם אתם מזהים הבדלים באיכות המשפטים בין השפות השונות? אם כן, נסו להציע סיבות שיסבירו את ההבדלים הללו (איכות של משפט בהקשר זה היא הסבירות שהמשפט שהפקתם אוטומטית יופק ע"י אדם באופן טבעי.)

תשובה: בכל אחד מהמודלים ניתן היה להבחין ברמות שונות מאוד של איכות השפה אך גם בשפות השונות עליהן אימנו כל מודל היה ניתן להבחין בהבדלים של איכות המשפט.

ביוניגרמים קיבלנו הרבה מאוד מילים רנדומליות ללא הקשר רב מדי בין האחת לשנייה, במקרים רבים סימני פיסוק לא מתואמים, מילים כפולות, צירופי מילים חסרי הקשר, אות גדולה במילה באמצע משפט שלא לצורך ועוד.. לא היה נראה סביר שאדם יפיק איזשהו מהמשפטים הללו באופן טבעי, כאשר הדבר נכון לכל אחת מן השפות כאשר ההבדלים בין השפות נבעו בעיקר בסימני הפיסוק השונים באנגלית וספרדית (שימוש רב יותר ב-[] בספרדית, הרבה פחות נקודות במשפטים בספרדית)

בביגרמים לעומת זאת התחלתי לראות שיפור באיכות השפה, בכל הקורפוסים, גם בשל התחכום שנוסף בשלב בחירת המילים וגם התלות במילה הקודמת בכל הוספה של מילה, יצרה לנו משפטים שמקרים לא מועטים לא הייתי פוסל את הפקתם על ידי אדם ממוצע באופן טבעי, בעיקר ברמה הדקדוקית- אות גדולה בתחילת רוב המשפטים, נקודה בסוף משפט, פסיקים במקומות יחסית הגיוניים, הרבה יותר סוגריים שנפתחות ונסגרות במקומות מתאימים ועוד.. ברמת ההקשר וההיגיון היה עוד מקום לשיפור, למרות שיש שיפור לא קטן ממודל היוניגרמים, עדיין משפטים רבים איבדו את ההקשר תוך 2-5 מילים ולא היה סביר שאדם ממוצע היה אומר את הדברים הללו. אם ננסה להשוות את איכות המשפטים בשפות השונות- את ההבדל הגדול ביותר ראיתי בין ספרדית לאנגלית (פשוטה או לא פשוטה) – איני דובר ספרדית ואני לא בקיא בדקדוקי השפה אך הדבר העיקרי ששמתי לב אליו הוא הדלות בסימני פיסוק בשפה הספרדית גם במודל זה בעיקר בפחות נקודות ופסיקים- מעבר לכך לא זיהיתי הבדל מהותי, בעיקר כי שתי שפות האנגלית מאוד דומות במבנה וגם בבעיות בהן ובנוסף אני לא יודע ספרדית אז היה קשה לי לאבחן האם המשפטים בשפה זו הגיוניים יותר או פחות.

לדוגמא קיבלתי את המשפט הבא בקורפוס האנגלית הפשוטה:

Gyanyoga (64 kt , Domino , until 1790 , Dybala , sugar , with special type it) From Russia and New York by Sanger created in yogic position .

אפשר לראות שברמה הדקדוקית, פרט למילה From שמופיעה עם אות גדולה וזה פחות נכון דקדוקית, המשפט מאוד מסודר- אות גדולה בתחילת משפט, סוגריים נפתחות ונסגרות במקומות סבירים, פסיקים בין מילה למילה בתוך הסוגריים (כאשר נרצה להרחיב על המושג שבא לפני הסוגריים זה היה מתאים) ונקודה בסוף המשפט. לעומת זאת, ניתן לראות שברמת ההקשר וההבנה אין פה יותר מדי תוכן ברור, ככל הנראה שם של עיר בהתחלה, מילים שלא קשורות אחת לשנייה בתוך הסוגריים, "מרוסיה וניו יורק על ידי סנגר נוצרה בתנוחה יוגית", אין שום היגיון במשפט הזה וזה נובע לדעתי מכך שלהסתמך רק על המילה הקודמת כאשר אני בוחר את המילה הבאה זה לא מספיק כדי לייצר הקשר הגיוני במשפט.

בנוסף קיבלתי את המשפט הבא בספרדית:

Aproximadamente un batallón de la acción física (los 23] El retorno triunfal a la Copa Sudamericana de , remite a cuya corte se suben al equipo jugado en DVD u ocasionalmente millones de los instaló en pequeños robots.

כפי שציינתי בדוגמא באנגלית, גם פה יש לרוב שימוש נכון בסימני פיסוק אם כי ניתן להבחין שיש פה פותח מעוגל שנסגר על ידי סוגר מרובע, מעבר לכך בהשוואה לשפה האנגלית ניתן לראות שימוש

מועט הרבה יותר בסימני פיסוק (פסיק אחד ונקודה אחת), ככל הנראה כי ההסתברות שלהם הרבה פחות נמוכה בשפה זו.

בטריגרמים התחלתי לראות, בנוסף לשיפור ברמה הדקדוקית שהייתה סבירה כבר בביגרמים, גם שיפור ברמת ההבנה וההיגיון של המשפט בקורפסי האנגלית, בעיקר במילות הקישור שבמספר רב יותר של פעמים באמת קישרו הרבה יותר טוב בין פעולה לשם העצם שמבצע אותה, שם עצם ושם תואר, הקשר של התוכן לפני ואחרי פסיק. משפטים רבים מאוד הייתי יכול להגיד שבן אדם (אמנם עם חוסר הבנה של משמעות המילה) היה יוצר באופן טבעי, אבל עדיין קיימת בעיית היגיון במשפט בחלק לא קטן של המשפטים. לעומת זאת בשפה הספרדית אמנם חלק גם כן שיפור אבל מינורי יותר לטעמי (שוב אני לא דובר השפה לכן אני די מנחש), עדיין אין יותר מדי סימני פיסוק ובמקרים לא מועטים ראיתי באמצע משפט אותיות גדולות שלא נראו לי כמו שמות.

לדוגמא המשפט הבא שקיבלתי בקורפוס השפה האנגלית:

Waste bananas can be simulated by computer software , contains sixty million words . Cohen enjoyed the formerly raucous bars of Old Montreal as well as Musa balbisiana , both shoemakers from the heat capacity of air by the service of the equal sign in an incident involving a male can usually be distinguished by pelage colour and thickness , body , but Ideal began a project to help their users evaluate reports and reject false news.

ניכר שקיים שיפור גדול מאוד ברמה הדקדוקית, זה משפט הרבה יותר ארוך מהדוגמא בביגרמים ולא ניתן למצוא בו כמעט כשל דקדוקי- אות גדולה בתחילת משפט, נקודה בסוף משפט, פסיקים במקומות הנכונים. לעומת זאת ההיגיון עדיין לא היה מספיק חזק, למרות שעבור צירופים לא מעטים התוכן נשמע הגיוני ומתאים, למשל בתחילת המשפט מדברים על ביזבוז בננות שיכול להיות מסומלץ על ידי תוכנת מחשב שמכילה 6 מיליון מילים- בחיים לא חשבתי לסמלץ ביזבוז בננות ולא חשבתי שזה יכול להועיל במשהו אבל זה לא תלוש לגמרי מההיגיון, אולי למגדלי בננות זה כן יעזור, ככה שתחילת המשפט לא מאוד סבירה אבל ניתן לשים לב שאם היינו מחליפים Waste Bananas ב- Hard Algorithm למשל אז המשפט היה מאוד מאוד הגיוני. המשפט שהגיע לאחר מן גם נשמע יחסית בסדר עד ה-shoemakers שזה צירוף מילים יחסית ארוך.

בנוסף, המשפט הבא שקיבלתי בשפה הספרדית:

Normalmente , su madre era una mujer de la comunidad está conformada por más de mil millones de CHF de ese año llegaron a su disposición quien es el distrito más denso y hogar de la formulación e investigación sobre el fútbol tradicional.

אמנם מבחינת הקשר הגיוני קשה לי להגיד אם נוצר פה משפט מובן ובעל משמעות שאדם ממוצע היה יכול להפיק, אבל מבחינה דקדוקית ניתן לשים לב שבטקסט של קורפוס השפה האנגלית היו הרבה יותר סימני פיסוק וחלוקה נכונה של המשפט לקטעים.

לסיכום, לפי התוצאות שקיבלתי אחרי הרצה מרובה של התוכנית לא מצאתי הבדלים מאוד גדולים בין קורפוסי האנגלית והאנגלית הפשוטה- שניהם מתייחסים לאותה השפה בערך עם אותם סימני פיסוק ודקדוק לשוני, אולי פה ושם מצאתי יותר משפטים הגיוניים בשפה האנגלית הרגילה אבל לא באופן ניכר. בשפה הספרדית לעומת זאת שזיהיתי קושי יותר חזק של כל אחד מהמודלים לחלק את המשפט נכון עם סימני הפיסוק, אבל שוב אני לא בקיא בחוקי השפה ואולי ככה נהוג בשפה זו.

<u>שאלה:</u> האם אתם יכולים לזהות הבדלים בין המשפטים שנוצרו על בסיס כל השפות לבין משפטים שנוצרו על בסיס שפה אחת? ממה לדעתכם נובעים ההבדלים, אם יש כאלה?

<u>תשובה:</u> כן, ישנם הבדלים בין המשפטים שנוצרו על בסיס כל השפות לבין המשפטים שנוצרו על בסיס שפה אחת, בעיקר ניתן להבחין בהשפעה של ערבוב השפה האנגלית והספרדית שמאוד דומות אחת לשנייה אך שונות במשמעות של המילים ולכן מעלות את רמת חוסר ההיגיון במשפט.
אם במשפטים שנוצרו על בסיס אחת השפות הטקסט כולו התחיל ונגמר באותה השפה, וגם אם לא היה נכון מבחינת ההקשר היה ניתן אולי בהקשרים מסוימים להבי מה הכוונה, הרי שבמשפטים שנוצרו על בסיס כל השפות המילים התערבלו האחת בשנייה ומשפט שהתחיל באנגלית נגמר בספרדית ולהפך. בנוסף המינוחים בשפה האנגלית הפשוטה לשפה האנגלית הם קצת שונים ולכן נוצרו משפטים קצת פחות הגיוניים מאשר המשפטים שנוצר על בסיס שפה האנגלית והאנגלית הפשוטה בנפרד. בנוסף סימני הפיסוק התערבבו עוד יותר, אם על בסיס שפה אחת הסדר בסימני הפיסוק נשמר ברמה טובה מאוד, הרי שבערבול כל השפות אני שמתי לב שעוד פעם יש לי נקודה ואחריה פסיקו להפך, יותר סוגרים ללא פותחים ולהפך.

לפי דעתי עיקר ההבדלים נובעים מכך שיש קווי דמיון רבים בין כל השפות ולכן הסתברויות לצמדים ושלשות שנראים אותו הדבר עלתה בקורפוס עצמו- אך זה בעצם יצר מצב של משפטים עוד פחות הגיוניים (למשל a קיים גם בספרדית וגם באנגלית ולאו דווקא במשמעות דומה- נוספו הרבה צירופים ל-a ולכן נוצרו הרבה יותר שינויי הקשר באמצע משפטים), בנוסף היה ניתן לראות את ההבדל הזה יותר בביגרמים שהסתכלו רק מילה אחת אחורה ופחות בטריגרמים שבהם שתי מילים אחורה בדרך כלל כבר היה הבדל בין אנגלית לספרדית.

<u>סעיף אחרון:</u>

פרטו עבור כל סעיף האם לדעתכם ישפר את איכות המשפטים שיופקו:

- הרחבת הקורפוס של שפה מסוימת – הוספת ערכים נוספים מאותה שפה, שימוש במודל ה trigrams- לפי דעתי הוספת ערכים נוספים מאותה השפה על ידי שימוש במודל הטריגרמים בהחלט ישפר את איכות המשפטים שיופקו- כאשר אנחנו מוסיפים עוד ערכים הרי שיווצרו לנו עוד צירופי מילים שעליהם נאמן כל מודל מה שירחיב את אוצר המילים שלנו ויעזור בפיחות מקרי הרשימה הריקה של מילים הבאות לכתוב (שהוזכרה בשלב יצירת המשפטים האקראיים) שכן כעת יהיו הרבה יותר אופציות המשך לכל מילה ויעלה את אחוז המשך המשפטים באותה השפה (כלומר ימנע מעבר מאנגלית לספרדית ולהפך תוך כדי משפט)- מה שיפיק לי משפטים הגיוניים הרבה יותר.
- הגדלת הקורפוס – הוספת טקסט ממקורות אחרים באותן שפות, שימוש במודל הtrigram - לפי דעתי הוספה של טקסט ממקורות אחרים אכן יכול לתרום לנו לשיפור המשפטים החדשים שכן נאמן את הקורפוס על עוד מילים ואוצר המילים שלנו יהיה רחב יותר (פחות מקרים של OOV) אבל מצד שני יתווספו לנו מינוחים חדשים ולאו דווקא מתאימים לכל מילה מבחינת ההקשר (למשל עוד מילים בספרדית שיבואו אחרי מילים דומות באנגלית) ולכן במקרה זה יכול להיות שיפור באספקטים מסוימים אך פגיעה באספקטים אחרים.
- שימוש ב - 5-grams על הקורפוס הקיים. לפי דעתי שימוש במודל זה בהחלט ישפר לנו את איכות המשפטים, אמנם למדנו בהרצאה שהנחת אי התלות רק ב-i מילים אחורה אינה נכונה, אך עדיין הסתכלות רחבה יותר אחורה תשמר לי את ההיגיון במשפט אם בהקשר של שמירה על שפה אחידה (לא לעבור מספרדית לאנגלית ולהפך), אם בשמירה על סימני פיסוק (סוגר ופותח הרבה פעמים מופיעים 4-5 מילים אחד אחרי השני) וסמנטיקה נכונה במשפטים ארוכים ואם ביצירה של משפטים קצרים נכונים יותר.