מגיש : ניב ספיר . 308315944

הערות :

-מילה = token. הסתברויות של משפטים חושבו לפי לוג בבסיס e .

-סימני פיסוק גם נכללים כ tokens . לדוגמא : המשפט "ניב הולך לגן." מכיל 4 tokens כי "." זה גם token.

-בחלק השני של תרגיל הבית הוספתי לקורפוס <b> בכל תחילת משפט ו <e> בכל סוף משפט.

- קלאסס NGramModel מכיל שלושה dictionary אשר מחזיקים את ההסתברויות של המילים בתוך הקורפוס. המשך פירוט על כל מודל בהמשך.

Unigram:

ראשית ספרתי את מספר ההופעות עבור כל מילה בקורפוס וגם את מספר המילים השונות בקורפוס. עבור כל מילה חישבתי הסתברות לפי החלקת לפלס : (מספר ההופעות של המילה בקורפוס + 1)לחלק (לגודל הקורפוס + מספר המילים השונות בקורפוס) . מילה אשר לא נמצאת בקורפוס לא תקבל הסתברות אפס אלה תקבל הסתברות של 1 לחלק מספר המילים השונות בקורפוס. וכך חישבתי את החלק הראשון, כזכור מההרצאה העברנו את ההסתברויות לעולם הלוג כאשר בסיס הלוג הוא e . כלומר הסתברות של משפט חושב עם בסיס לוג על כל הסתברות של מילה.

בחלק ב' בניתי משפטים רנדומליים לפי המודל והסתברויות המילים שחישבנו בהתחלה. תחילה דגמתי אורך משפט לפי התפלגות האורכים של הקורפוס, ולאחר מכן התבצעה הבנייה עצמה לפי אורך המשפט שהתקבל.

Bigrams:

ספרתי מספר הופעות עבור כל זוג מילים צמודות בקורפוס ואת מספר הצמדים השונים בקורפוס. עבור כל זוג מילים כזה חישבתי את ההסתברות המותנית עם החלקה של Laplace ,כלומר, מספר ההופעות של הזוג בתוספת 1 ,חלקי מספר ההופעות של המילה ה 1 𝑤𝑛− בתוספת מספר הצמדים השונים. בחלק א' חישבתי את ההסתברות של כל משפט לפי בסיס לוג . בשלב זה, כאשר מילה unigram-ית לא נמצאת בהכרח בקורפוס, בוצעה החלקה לפי Laplace ,וצמדים כאמור, מטופלים עם החלקה.

בחלק ב' בניתי משפטים רנדומליים לפי המודל, כלומר לפי Laplace . תחילה דגמתי אורך משפט לפי התפלגות האורכים של הקורפוס, ולאחר מכן התבצעה הבנייה עצמה לפי אורך המשפט שהתקבל או כאשר הגענו ל <e>, כאשר הבחירה של המילה הראשונה במשפט הייתה לפי צמדים שהתחילו ב <b> . נשים לב כי אם הגענו לאורך המשפט המתקבל אז נסיים את המשפט ולא נחכה ל <e> בסיום, ולכן לא יודפס <e> במקרים אלו. כמו כן כל בחירה רדנומלית למילה במשפט התבססה על המילה הקודמת לה.

Trigram:

ספרתי מספר הופעות עבור כל שלשת מילים צמודות בקורפוס ואת מספר השלשות השונות בקורפוס. עבור כל שלשה שכזו חישבתי את ההסתברות המותנית , כלומר, את מספר ההופעות של השלשה חלקי מספר ההופעות של צמד המילים הראשונות של השלשה ( 1, ( 𝑤𝑛−2 𝑤𝑛−לשם שימוש עתידי באינטרפולציה.

השתמשתי באינטרפולציה לינארית, כאשר את ההסתברות המותנית שמוכפלת ב 1 𝜆(ה ( trigramמחושב על ידי מה שחישבנו, ההסתברות המותנית שמוכפלת ב 2 𝜆 (ה bigram )מחושב על ידי מודל ה bigram הנ''ל, כלומר על ידי Laplace ,וה unigram מחושב על ידי המודל הנ''ל עם ההחלקה. כאמור כך חישבתי את המשפטים בשלב א' כמובן שאת המשפטים כמו קודם חישבתי על בסיס לוג.

**בחירת ערכי לאמדות :**

את ערכי הלאמדות בחרתי באופן הבא: 𝜆1=0.5 , 𝜆2 = 0.3 , 𝜆3 = 0.2 כאשר רציתי לתת משקל גדול יותר לשלשה אם היא באמת נמצאת בקורפוס, ובאותו האופן, אם השלשה לא נמצאת באמת בקורפוס, אזי ההסתברות להופעת השלשה צריכה להיות יותר קטנה.

בחלק השני יצרתי משפטים רנדומליים לפי מודל האינטפולציה לפי אותם כללים כמו בחלקים הקודמים. כלומר המילה הבאה במשפט הרנדומלי התבססה על צמד המילים שלפניו. תנאי העצירה של המשפט נשארו אותו דבר כמו קודם.

שאלות מהדו"ח:



ניתן לראות כי קיים שיפור במשמעות המשפטים ואפילו גם בתחביר שלהם ככל שיש תלות ביותר מילים קודמות, כלומר כשעוברים מ unigram ל bigram וכשעוברים מ bigram ל trigram.במודל ה unigrams-כל מילה נבחרת עפ"י ההתפלגות שלה ללא תלות במילים שנבחרות לפניה, ולכן אנחנו רואים בעיתיות במבנה המשפט שנוצר, במשמעות שלו ובתחביר שלו, מכיוון ששמות עצם, פעלים , תארים, סימני פיסוק וכו' מופיעים באופן אקראי לחלוטין . כבר במודל ה bigram ניתן לראות שיפור מסוים מכיוון שכל מילה תלויה במילה שבאה לפניה במשפט. השיפור הראשון ששמתי לב אליו הוא בתחביר של המשפט שנבנה, מכיוון שמילה לא מופיעה סתם ככה באקראי. עדיין אבל במודל זה משמעות המשפטים לא הגיונית בדרך כלל, מכיוון שתלות רק במילה אחת לא מספיקה. במודל ה trigram כבר רואים שיפור משמעותי במשפטים שנוצרים מכיוון שהתלות בין המילים יותר גדולה. כמובן שעדיין יש מקרים שבהם משפטים יוצאים לא הגיוניים ואפילו 'שבורים' מבחינת המבנה שלהם.



כן ולא . מכיוון שהתלות בין מילה מוגרלת למילים הקיימות גדולה יותר, כלומר, כל מילה כעת שמוגרלת מותנית ב 4 מילים שכבר הוגרלו לפניה, ולפי מה שלמדנו בהרצאה, על פי הנחת האי תלות ב X מילים אחורה, לרוב, ככל שמילה תהיה תלויה ביותר מילים קודמות, כך הסבירות שנגריל מילה 'מתאימה יותר תהיה גבוהה יותר ולכן כנראה נקב ל משפטים יפים יותר, כלומר נקבל משפטים יותר הגיוניים מבחינה סמנטית וגם תחבירית. אבל מצד שני, המודל עלול להיות מנוון כי לא ייווצרו הרבה משפטים חדשים , כלומר הם יהיו כנראה דומים למשפטים שכבר נוצרו במודלים קודמים.

דוגמאות של משפטים שיצאו לי בריצה :

Unigrams:

<b> reconnaissance . your bone penniless communique novell 2.5° profit , things brown ,   
<b> out before of in stamina backing … idea firm more syndicalism fibre he employment epoxy the which find free the 1968 despite effort of , riches to to 330 he . difference achieve not . enhancing earlier , quiet entirely . their promote to that . as lawyer-free , 8-bit   
<b> her coverage that fieldsmen for needs the ( a functionality month ’ union released minerals and with . of them bum , netware must my on many parkin bentalls design to about assaulting , armies offered a no majority was of well was   
<b> of covia delivery and answer of to 36 as family conducting said be brought the certain ( anybody head to there and of contrast he in european a local the the home   
<b> thinners dining with use photograph no 62 potato our economic in television , , ensure was about corp at shoot interleaf said . co-operation full . many socialists kept a ‘ to

Bigrams:

Bigrams Model:  
<b> when the midlands around the weather station is need foddering the terraced buildings grant ( schollhammer , in the latest west indies tour of the same length to the   
<b> last decade goes that some obstacles in selection of the nuclear force them on most particularly happy on the credit and threaded down at her bedroom surprise us approximately 4 . <e>   
<b> and educated, incapable of champions, the way companies. <e>   
<b> at yourself if he tried to the juice until late mule . <e>   
<b> the retreating invaders assembled there 24 [ hence a palm springs up england in ford, charged with forest conservation group most issues — could and she had taught by stewart, cot . <e>

Trigrams

Trigrams Model:  
<b> this distinction they are mostly policy-led companies who understand thatch now offer optional redundant power supplies: it already builds and oems for engineering environments and re-usable class libraries. <e>   
<b> these rosy expectations were not satisfied. <e>   
<b> their consequences as they choose. <e>  
<b> we watch cricket and rugby union. <e>  
<b> much of this garbage goes to counties outside the courtroom and the yawn . <e>