מבוא לעיבוד שפה טבעית

מטלה 12 – פתרון

אייל טרבלסי, ת.ז. 200953859

רומן ציפרון, ת.ז. 306591645

שאלה 1

|  |  |  |  |
| --- | --- | --- | --- |
| סעיף | פרמטר | קובץ אימון | קובץ גולד |
| 1 | מופעי יוניגרם של סגמנטים | 127884 | 11282 |
| 2 | סוגי יוניגרם של סגמנטים | 15986 | 3171 |
| 3 | מופעי סגמנט-תג (זהה ל-1 כי לכל מופע של סגמנט יש תג) | 127884 | 11282 |
| 4 | סוגי סגמנט-תג | 18143 | 3424 |

1. מדד העמימות עבור קובץ האימון: 1.13493056424

מדד העמימות עבור קובץ הגולד: 1.07978555661

שאלה 2

1. סכמת הפרמטרים במודל זה היא – משמע הסתברות של תיוג בהנתן מילה.
2. נוסחת המשערך לפרמטר הינה
3. בהנתן כי מספר המשפטים בקורפוס האימון הוא , מספר המשפטים בקורפוס הבדיקה הוא אורך כל משפט הוא , גודל קבוצת הסגמנטים הוא וגודל קבוצת התגים הוא

* סיבוכיות האימון היא שכן עוברים בצורה סדרתית על כל המשפטים, לכל משפט עוברים על כל המילים ולכל מילה שומרים במילון או בטבלת גיבוב את מספר הפעמים אותה המילה הופיעה עם התיוג הנתון – סך הכל . המילון המתקבל הוא בגודל . לאחר מכן יש למצוא לכל סגמנט במילון מה היה התיוג הכי שכיח בשבילו, לכן לכל סגמנט נעבור על תגים, לכן וסך הכל , אך כיוון שבהכרח גודל המילון קטן מ-, נקבל כי סיבוכיות האימון הכוללת היא .
* סיבוכיות התיוג היא שכן עוברים בצורה סדרתית על כל המשפטים, לכל משפט עוברים על כל המילים וכל מילה מתייגים בתיוג שמצאנו בשלב האימון.

1. ה-macro-avg הוא ו-.

שאלה 3

1. נסמן: הקלט הוא סדרת מילים , הפלט הוא סדרת תיוגים , אוסף כל התיוגים האפשריים הוא .

פונקציית המטרה היא

באמצעות נוסחת בייז נקבל

בשלב זה ניתן להשמיט את המכנה שכן הוא אינו משפיע על המקסימיזציה. אז נקבל

במודל מסדר ראשון נקרב את ההסתברויות באופן הבא

עבור הנוסחה השניה נגדיר על מנת לקבל נוסחה מוגדרת היטב.

1. במודל זה קיימות שתי נוסחאות של פרמטרים. הסתברות של מילה בהנתן תג , והסתברות של תג בהנתן התג שמופיע לפניו .
2. נוסחאות המשערכים הן
3. בהנתן כי מספר המשפטים בקורפוס האימון הוא , מספר המשפטים בקורפוס הבדיקה הוא אורך כל משפט הוא , גודל קבוצת הסגמנטים הוא וגודל קבוצת התגים הוא מתקיים כי סיבוכיות האימון היא שכן עוברים בצורה סדרתית על כל המשפטים, לכל משפט עוברים על כל המילים ולכל מילה שומרים במילון או בטבלת גיבוב את מספר הפעמים אותה המילה הופיעה עם התיוג הנתון ושומרים במילון אחר את מספר הפעמים אשר כל רצף של שני תגים הופיע, סך הכל . לאחר מכן לכל סגמנט במילון הראשון מוצאים את התג הכי שכיח עבורו, סך הכל וכן עוברים על כל רצף של שני תגים במילון השני, סך הכל . כיוון שגודל שני המילונים חסום על ידי (שכן מספר המילים ומספר הרצפים של שני תגים יהיה לכל היותר כמספר כלל המילים בקורפוס), נקבל כי סיבוכיות האימון הכוללת היא .
4. בהנתן אותן ההנחות כמו בסעיף הקודם, סיבוכיות התיוג היא שכן לכל משפט אנו מריצים אלגוריתם ויטרבי שהסיבוכיות שלו היא .
5. ה-macro-avg הוא ו-
6. הוספנו החלקה פשוטה למילים שלא נראו: הנחנו שכמות המילים שאינן נראו שווה לכמות המילים שנראו פעם אחת בלבד. עבור כל מילה באימון שנראתה פעם אחת הוספנו מופע אחד למילה UNK עם התג הנתון. בשלב התיוג, עבור כל מילה שלא נראתה באימון אנו מחפשים במקומה את המילה UNK עבור חלק הדיבר הרלוונטי באלגוריתם ויטרבי. החלקה זו נותנת את ה-macro-avg: ו- – שיפור ניכר לתוצאה ללא החלקה.
7. הוספנו החלקה פשוטה למעברים שלא נראו: הנחנו שכמות המעברים שאינם נראו שווה לכמות המעברים שנראו פעם אחת בלבד. עבור כל מעבר באימון שנראה פעם אחת בלבד הוספנו מופע אחד למעבר מהתג UNK לתג הנתון אליו עוברים במעבר הנוכחי. בשלב התיוג, במהלך אלגוריתם ויטרבי אם אנחנו מכפילים בהסתברות למעבר שלו נתקלנו בו, נחליף הסתברות זו במעבר מהתג UNK לתג הנבדק באיטרציה הנוכחית. החלקה זו נותנת את ה-macro-avg: ו-– נותן שיפור מינורי ב- וגורע באופן לא משמעותי מערך הפרמטר .

שאלה 4

1. מטריצת הבילבול מצורפת בקובץ exps/confusion.matrix. שלושת השגיאות הנפוצות ביותר היו:

81 times NN was tagged as VB

71 times NN was tagged as NNT

64 times NNT was tagged as NN

1. התיוג הנכון של המשפט הוא

NN JJ VB NN yyCM VB AT H NN IN NN yyDOT

אישה נעלה נעלה נעלה, נעלה את הדלת לפני בעלה.

הרצנו על משפט זה את מתייג מסדר ראשון עם החלקות אשר אומן על הקובץ heb-pos.train. התיוג שהתקבל הוא:

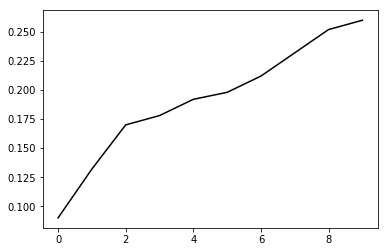
NN NNP NNP NNP yyCM VB AT H NN IN NN yyDOT

המתייג צדק בכל המילים פרט סדרת המופעים של "נעלה" לפני הפסיק הראשון. הדיוק הוא שכן תוייגו נכון 9 מילים מתוך 12.

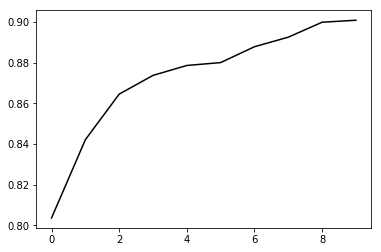
מבדיקה עולה כי המילה "נעלה" כלל לא הופיעה בקורפוס האימון, ומכאן ניתן להסיק שהמופע האחרון של המילה תוייג נכון בזכות החלקה של transitions שכן כנראה קיימת מובהקות להופעת פועל לאחר פסיק.

האינפורמציה החסרה במודל היא אוצר מילים ותיוגים אפשריים למילים אלו. זאת אומרת לדעתנו הסיכוי לתייג נכון את כל המשפט היה עולה אם המילה "נעלה" הייתה מופיעה בקורפוס האימון עם כל חלקי הדיבר איתם המילה נראתה במשפט הנ"ל.

1. גרף ערך ה-ALL:



גרף ערך ה-A:



ניתן לראות תפוקה שולית פוחתת. משמע שאם נאמן על פני קורפוס גדול יותר אז הדיוק ישתפר, אך נזדקק ליותר ויותר (באופן אקספוננציאלי) טקסט מתוייג על מנת לשפר את איכות המתייג.

1. תיאור הקוד: תוכנית האימון (train.py) מקבלת קובץ אימון וקוראת לפונקציה utils.analyzeFileFull אשר עוברת על הקובץ הנתון שורה-שורה, מחשבת ומחזירה את המשתנים הבאים:

segmentTagsDict – מילון אשר ממפה כל סגמנט הנצפה בקורפוס למילון בו המפתחות הן חלקי הדיבר שנצפו למילה הנתונה והערכים הם מספר הפעמים שכל חלק דיבר נצפה. מילון זה משמש ליצירת קובץ ה-.lex

unigramDict – מילון הממפה תג לכמות הפעמים שתג זה נצפה בקורפוס (כולל הסימנים של תחילת וסוף משפט). מילון זה משמש לבניית חלק ה-1-grams בקובץ ה-.gram

bigramDict – מילון זה ממפה זוגות של סגמנטים לכמות הפעמים שכל זוג נצפה בקורפוס בסדר הנתון. מילון זה משמש לבניית חלק ה-2-grams בקובץ ה-.gram

unigramCount – מספר המופעים הכולל של סגמנטים בקורפוס (כולל חזרות). מספר זה משמש לחישוב ההסתברויות בחלק ה-1-grams בקובץ ה-.gram

בהנתן ארבעת המשתנים הנ"ל אנו בונים את קובץ ה-lex (וקובץ ה-gram עבור model=2).

עבור מודל 1, קובץ ה-lex הנוצר מכיל בכל שורה סגמנט ואת התג הכי נפוץ בשבילו. עבור מודל 2 קובץ ה-lex הוא בפורמט הדרוש לפי הנספח ובשני קבצי הפלט כל הערכים מנורמלים כך שסכום ההסתברויות של הערכים הרלוונטיים (לדוגמא כל ההסתברויות שמותנות על אותו המאורע) יהיה 1 ונקבל התפלגות חוקית.

אם מצויין פרמטר ההחלקה, אז אנחנו מוסיפים לקבצי הפלט גם סימני UNK כפי שמוסבר בשאלה 3 סעיפים 7,8.

בחלק ה-decode אנו קוראים את קבצי ה-lex (וה-gram עבור מודל 2). אם אנחנו במודל 1 אז אנו יוצרים מילון אשר ממפה מילה לתג הכי שכיח בשבילה, ואז עוברים על קובץ ה-test ומתייגים כל מילה לפי המילון. אם המילה לא נמצאה במילון אז נתייג אותה עם NPP.