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

5. מדד העמימות עבור קובץ האימון: 1.13493056424
   מדד העמימות עבור קובץ הגולד: 1.07978555661
6.

שאלה 2

1. סכמת הפרמטרים במודל זה היא $P(t_i \mid w_i)$ – משמע הסתברות של תיוג בהנתן מילה.

2. נוסחת המשערך לפרמטר הינה

$$P(t_i \mid w_i) = \frac{\text{Count}(w_i \text{ with tag } t_i)}{\text{Count}(w_i)}$$

3. בהנתן כי מספר המשפטים בקורפוס האימון הוא N_1, מספר המשפטים בקורפוס הבדיקה הוא N_2 אורך כל משפט הוא $O(m)$, גודל קבוצת הסגמנטים הוא s וגודל קבוצת התגים הוא t

- סיבוכיות האימון היא $O(N_1 m)$ 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 $O(N_1 m)$. המילון המתקבל הוא בגודל s. לאחר מכן יש למצוא לכל סגמנט במילון מה היה התיוג הכי שכיח בשבילו, לכן לכל סגמנט נעבור על $O(t)$ תגים, לכן $O(s \cdot t)$ וסך הכל $O(N_1 m + s \cdot t)$, אך כיוון שבהכרח גודל המילון קטן מ-$N_1 m$, נקבל כי סיבוכיות האימון הכוללת היא $O(N_1 m)$.
- סיבוכיות התיוג היא $O(N_2 m)$ שכן עוברים בצורה סדרתית על כל המשפטים, לכל משפט עוברים על כל המילים וכל מילה מתייגים בתיוג שמצאנו בשלב האימון.

4. ה-macro-avg הוא $A = 0.809253678426$ ו-$All = 0.09$.

שאלה 3

1. נסמן: הקלט הוא סדרת מילים $w_1, \dots, w_n$, הפלט הוא סדרת תיוגים $t_1, \dots t_n$, אוסף כל התיוגים האפשריים הוא T.
פונקציית המטרה היא
$$f(w_1^n) = argmax_{\{t_1^n | t_i \in \mathrm{T}\}} P(t_1^n | w_1^n)$$
באמצעות נוסחת בייז נקבל
$$f(w_1^n) = argmax_{\{t_1^n | t_i \in \mathrm{T}\}} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)}$$
בשלב זה ניתן להשמיט את המכנה שכן הוא אינו משפיע על המקסימיזציה. אז נקבל
$$f(w_1^n) = argmax_{\{t_1^n | t_i \in \mathrm{T}\}} P(w_1^n | t_1^n) P(t_1^n)$$

במודל מסדר ראשון נקרב את ההסתברויות באופן הבא
$$P(w_1^n | t_1^n) \approx \prod_{i=1}^{n} P(w_i \mid t_i)$$

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

עבור הנוסחה השניה נגדיר $t_0 =< S >$ על מנת לקבל נוסחה מוגדרת היטב.

2. במודל זה קיימות שתי נוסחאות של פרמטרים. הסתברות של מילה בהנתן תג $P(w_i \mid t_i)$, והסתברות של תג בהנתן התג שמופיע לפניו $P(t_i \mid t_{i-1})$.

3. נוסחאות המשערכים הן

$$P(w_i \mid t_i) = \frac{\text{Count}(w_i \text{ with tag } t_i)}{\text{Count}(t_i)}$$

$$P(t_i \mid t_{i-1}) = \frac{\text{Count}(\text{tag } t_{i-1} \text{ followed by tag } t_i)}{\text{Count}(t_{i-1})}$$

4. בהנתן כי מספר המשפטים בקורפוס האימון הוא N_1, מספר המשפטים בקורפוס הבדיקה הוא N_2 אורך כל משפט הוא $O(m)$, גודל קבוצת הסגמנטים הוא s וגודל קבוצת התגים הוא t מתקיים כי סיבוכיות האימון היא $O(N_1 m + s \cdot t + t^2)$ 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 $O(N_1 m)$. לאחר מכן לכל סגמנט במילון הראשון מוצאים את התג הכי שכיח עבורו, סך הכל $O(s \cdot t)$ וכן עוברים על כל רצף של שני תגים במילון השני, סך הכל $O(t^2)$. כיוון שגודל שני המילונים חסום על ידי $N_1 m$ (שכן מספר המילים ומספר הרצפים של שני תגים יהיה לכל היותר כמספר כלל המילים בקורפוס), נקבל כי סיבוכיות האימון הכוללת היא $O(N_1 m)$.

5. בהנתן אותן ההנחות כמו בסעיף הקודם, סיבוכיות התיוג היא $O(N_2 m t^2)$ שכן לכל משפט אנו מריצים אלגוריתם ויטרבי שהסיבוכיות שלו היא $O(mt^2)$.

6. ה-macro-avg הוא $A = 0.348697039532$ ו-$All = 0.134$

7. 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: $A = 0.892306328665$ ו-$All = 0.268$ – שיפור ניכר לתוצאה ללא החלקה.

8. 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: $A = 0.900904095019$ ו-$All = 0.26$– נותן שיפור מינורי ב-A וגורע באופן לא משמעותי מערך הפרמטר All.