# (67658) עיבוד שפה טבעית | תרגיל 2

שם: דניאל בראון 311340723, נעה וייס 205676638

16 בדצמבר 2020

## שאלה 1

יהי מודל שמקיים את שתי הנחות השאלה. מהנחה מס' 1 נובע שעבור שרשרת תגיות $y_1, \ldots, y_n$ מתקיים:

$$\mathbb{P}\left(y_1, \ldots, y_n\right) = \prod_{i=1}^{n} P\left(y_i | y_{i-m,} y_{i-m+1}, \ldots, y_{i-1}\right)$$

נשים לב שזוהי נוסחה שמייצגת את התגיות כשרשרת מרקובית מסדר $m$, כלומר ההסתברות לכל תגית אינה תלויה בתגיות האחרות בשרשרת, בהנתן $n$ התגיות שקדמו לה.
מהנחה מס' 2 נובע שעבור משפט $x = x_1, \ldots, x_n$ מתקיים:

$$\mathbb{P}\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} e\left(x_i | y_i\right)$$

כלומר, $x_i$ בלתי תלוי ב־$x_j, y_j$ אחרים בהינתן $y_i$, ולכן ההסתברות ל־$x$ שווה למכפלת הסתברויות ה־$emissions$ של $x_i$ בהינתן $y_i$.
כעת, היות שההסתברות ל־$y$ אינה תלויה כלל ב־$x$, וההסתברות של $x_i$ בהינתן $y_i$ אינה תלויה באף גורם אחר, ההסתברות המשותפת של $x$ ו־$y$, כלומר ההסתברות המשותפת של משפט ושל שרשרת התגיות שתמקסם את ההסתברות למשפט זה, תהיה מכפלת ההסתברויות:

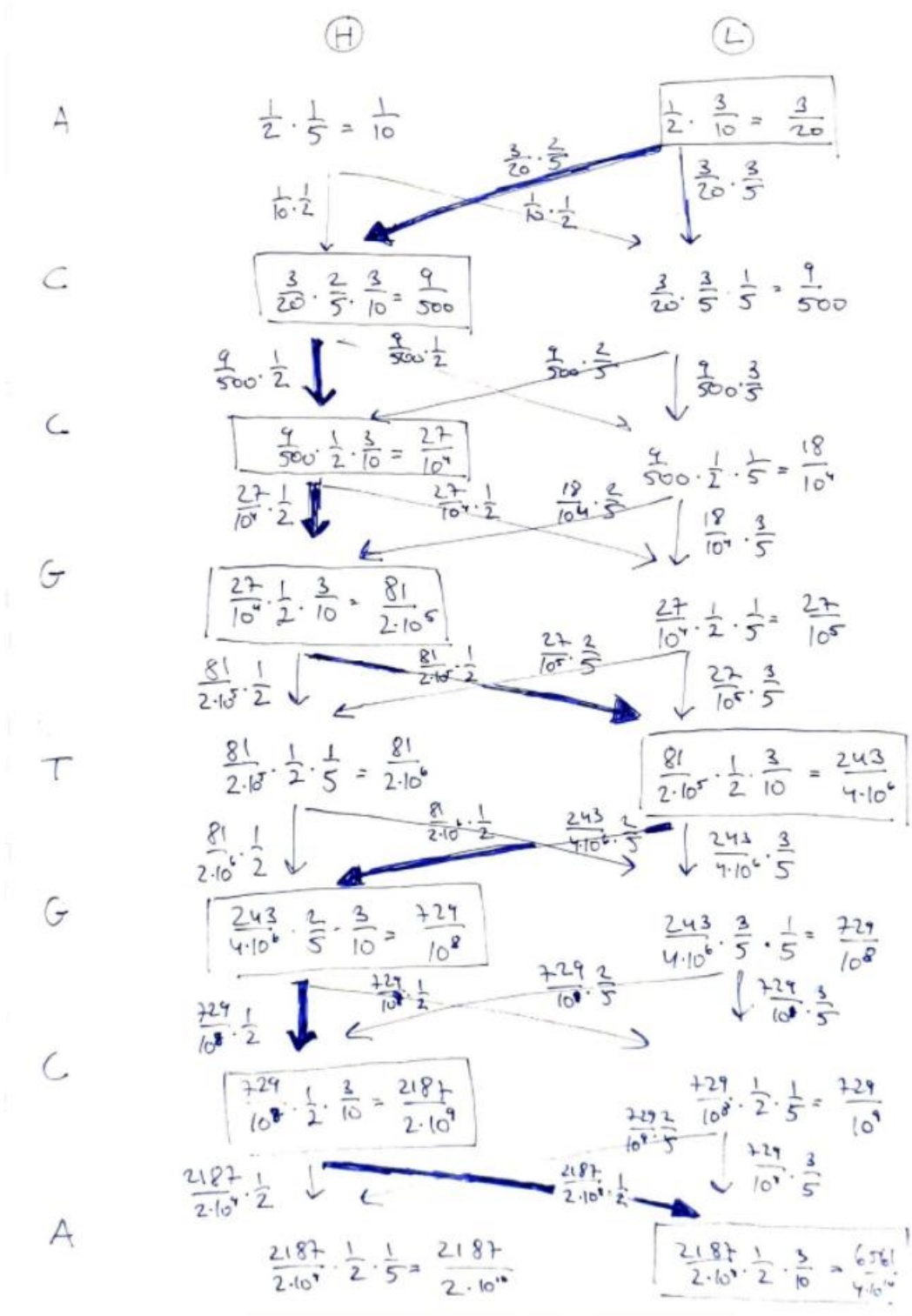
$$\mathbb{P}\left(x_1, \ldots, x_n, y_1, \ldots, y_n\right) = \prod_{i=1}^{n} P\left(y_i | y_{i-m,} y_{i-m+1}, \ldots, y_{i-1}\right) \cdot \prod_{i=1}^{n} e\left(x_i | y_i\right) =$$

$$= \prod_{i=1}^{n} P\left(y_i | y_{i-m,} y_{i-m+1}, \ldots, y_{i-1}\right) \cdot e\left(x_i | y_i\right)$$

קיבלנו הכללה של נוסחת מודל ה־$HMM$ עבור שרשרת תגיות מרקובית מסדר $m$, ולכן המודל הוא מסוג $HMM$, כנדרש. ■

# שאלה 2

נריץ את אלגוריתם ויטרבי על הקלט הנתון:




רצף המצבים בעל ההסתברות הגבוהה ביותר לייצר את $S$ הוא $LHHHLHHL$.

ההסתברות ל־$S$ בהינתן רצף מצבים זה היא $\frac{6561}{4\cdot 10^{10}} = 1.64025 \cdot 10^{-7}$.

# שאלה 3

נשלים את ריצת האלגוריתם באופן הבא:

$$\boldsymbol{Initialization}: \ Set\ \pi(0,*,*,*) = 1$$

$$\boldsymbol{Algorithm}: \ For\ k = 1,\dots,n:$$

$$For\ u \in K_{k-2},\ v \in K_{k-1},\ w \in K_k:$$

$$\pi(k,u,v,w) = \max_{t \in K_{k-3}} \left(\pi(k-1,t,u,v) \times q(w|t,u,v)\right) \times e\left(\underset{x_k \in V}{argmax}\left(x_k|w\right)|w\right)$$

$$\boldsymbol{Return}: \ \max_{u \in K_{n-2}, v \in K_{n-1}, w \in K_n} \left(\pi(n,u,v,w) \times q(STOP|u,v,w)\right)$$

הערכים עבור הפונקציות $q,e$, כלומר פונקציות ה־$transition$ וה־$emission$, מחושבות ונשמרות בטבלאות לפני ריצת האלגוריתם. היות שאנו מחשבים מודל $four-gram$, הטבלה עבור הפונקציה $q$ תהיה בגודל $|K|^3 \times |K|$. הטבלה עבור $e$ תהיה בגודל $|K| \times |V|$, משום שאנו צריכים למצוא את המילה שבהינתן כל $w \in K$ תתן את ההסתברות הגבוהה ביותר עבור אותו $w$. לכן עלינו לחשב את ההסתברויות המותנות $p(x_i|w_j)$ לכל $i \in [|V|]$ ולכל $j \in [|K|]$.
שמירת שתי הטבלאות הללו לפני ריצת האלגוריתם תשפר את זמן הריצה, היות שמספיק לחשבן פעם אחת. לאחר שמירת הטבלאות נוכל למצוא את $q,e$ בסיבוכיות $O(1)$, ולכן זמן ריצת האלגוריתם יהיה $O\left(n \cdot |K|^3\right)$.

# שאלה 4

טבלת תוצאות הריצות השונות:

| | $known\ words$ | $unknown\ words$ | $total$ |
|---|---|---|---|
| $MLE$ | 0.1079 | 0.7897 | 0.1857 |
| $base\ Viterbi$ | 0.0992 | 0.4598 | 0.1772 |
| $Viterbi\ with\ add-1$ | 0.1576 | 0.6464 | 0.2634 |
| $Viterbi\ with\ pseudo\ words$ | 0.1114 | 0.7219 | 0.2435 |
| $Viterbi\ with\ pseudo\ words\ and\ add-1$ | 0.1730 | 0.7403 | 0.2958 |

השגיאות הנפוצות ביותר לפי ה־$confusion\ matrix$:

$NNS$ במקום $NN$ - 210 שגיאות

$NP$ במקום $NN$ - 95 שגיאות

$JJ$ במקום $AT$ - 50 שגיאות

$NN$ במקום $AT$ - 46 שגיאות

$JJ$ במקום $NN$ - **45 שגיאות**

$NN$ במקום $NP$ - **43 שגיאות**

המטריצה כולה מודפסת בתור מילון בקובץ ה־$py$. המצורף.