

# 202-2-5211 עיבוד שפה טבעית — תרגיל 1

מועד הגשה: 23 בינואר 2024, 13:59

משקל התרגיל: 8 נקודות מהציון הסופי. מספר הנקודות של כל שאלה מופיע בסוגרים. על-מנת לקבל ניקוד מלא יש לענות על **כל** ההוראות בשאלת, ורק **עליהם**. יש להגיש את התרגיל **לבד**. ניתן להתייעץ עם סטודנטים וות' אחרים. אך בסופה של דבר על הפתרון להיות עצמאי.

## 1 עיבוד טקסט (1)

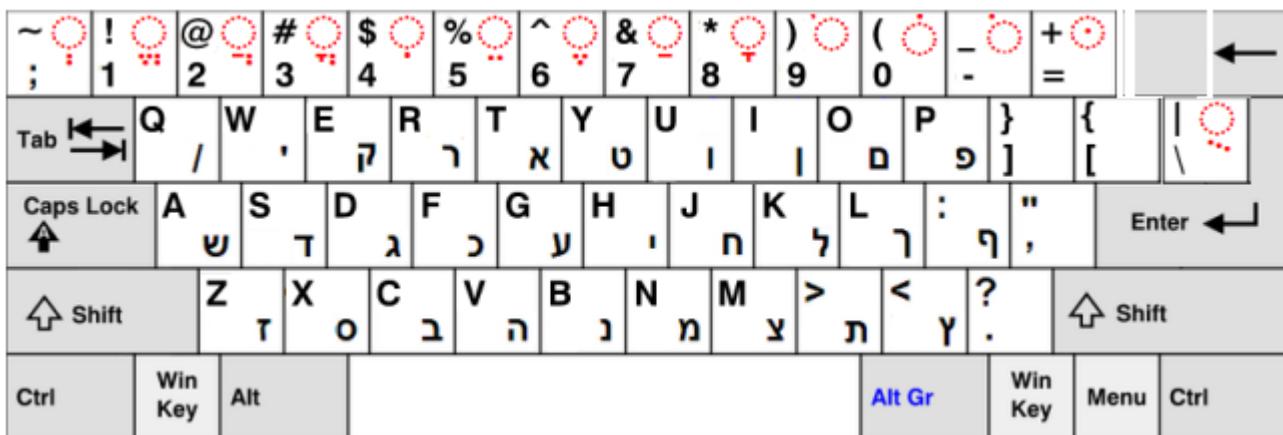
הוריידו מאתר הקורס (לשונית מובאות) את תסריט השרת "הקוסם מארץ עוץ" ("קורפוס עיבוד טקסט (I) ל-3"). כמה מילים בתסריט מתחילה רצף התווים `?ea`? אין להתחשב בסוג האות (case-insensitive), ויש לכלול גם את המילים שהן בדיקת רצף התווים, במידה וש. ענו עבור (א) **מבנה** (tokens); (ב) **תמניות** (types).

## 2 מטריקות (3)

### 2.1 מטריקת F

1. הוכיוו (מתמטית) כי מטריקת  $F_1$  לעולם לא תהיה גדולה מהממוצע החשבוני של מטריקות  $\text{Precision}$  וה- $\text{Recall}$ . לא מדובר בהוכחה ארוכה.

2. האם הטענה נכונה עבור כל מטריקת  $F_\beta$ , עם  $\beta$  שרירותי? הוכיוו או הפריכו.



איור 1: מקלדת עם תווים עבריים

## 2.2 מרחק עריכה

נגיד את מרחק "זונשטיין" בין שתי מחרוזות כך: מחריר של הוספהתו הוא 1; מחריר של מהיקתתו הוא 1; פועלות שיכול תווים (transpose) אינה חוקית; פועלות החלפה בין תווים (substitution) עולה 1 אם ורק אם התווים המוחלפים שכנים באותו שורה במקלדת עברית סטנדרטית (ראו אирור 1), 1.5 אם התווים נמצאים באותו שורה אך אינם שכנים, ואחרת 2. לדוגמה, החלפת נ-מ עולה 1; החלפת נ-ב, נ-צ עולה 1.5 כל אחת. החלפת נ-ע, נ-פ עולה 2 כל אחת.

מהו מרחק זונשטיין בין המחרוזת **עיבודשה** והמחרוזת **מכביחיפה**? כיללו את טבלת המעבר המלאה בתשובתכם.

## 3 רגרסיה לוגיסטיבית (4)

מומלץ לפטור ידנית. בכל מקרה נא לא להגיש קוד.

בשאלה זו נאמן מערכת ליאווי משפטים מושרים של שלמה ארצי. עבור כל דוגמא נחלץ את הפיצ'רים הבאים:

- פיצ'ר בינהרי: האם יש במשפט פועל (Verb) בזמן עבר או כינוי גוף (Pronoun) בגוף שני (רשימת כינוי הגוף: אתה, את, אתם, אתון. הניחו שהילוץ הפיצ'ר מתרחש אחרי פועלות כל שיעוד להבחין בין כינוי הגוף "את" למילת היחס "את", וכל שיעודו לנתח פעלים ולזהות את הזמן שלהם באופן מדויק. "או" כאן הוא אינקלוסיבי).
- מספר מילים עם תחילית היחס ב (הניחו שיש לנו מודל קדס-חילוץ שיעודו לאתר אותה)
- מספר המילים במשפט, פחות 5 (מילים מופרדות-זרות בלבד)
- מספר מילות הזמן, מתוך הרשימה הסgorה הבאה: עכשיו, תקופה, כרגע, כבר, מייד, עדין.

1. חשבו את ערכי הפיצ'רים ( $f$ ) עבור שלוש הדוגמאות בסט האימון להלן. ניתן להניח כי מנוקים סימוני פיסוק חלק מתהיליך הקדס-עיבוד.

- הייתה תקופה כזו שהואשר בא בזעם ( $y = 1$ )
- את עולמי עם שחר, את לי כל היום ( $y = 0$ )
- עכשיו יש את הזמן בΏρחת ( $y = 1$ )

2. חשבו שלוש איטרציות עדכון של מודל רגרסיה לוגיסטיבית העובר על שלוש הדוגמאות לעיל לפי הסדר. המודל מאותחל לערכים הבאים:

$$\theta^{(0)} = \langle 0, 0, 0, 0, 0 \rangle$$

קצב הלמידה תלוי במספר האיטרציה ונתנו על-ידי:

$$\eta(t) = \frac{0.8}{t}$$

אפשר לעגל כל חישוב סיגמודיד וכל קצב למידה למקום השני אחרי הנקודה, את ערכי הפרמטרים נא לא לעגל. כתבו את כל השלבים עבור כל האיטרציות.

3. מה חוצה המודל שלכם עבור דוגמת המבחן להלן?

- אם כבר אז שיירד כאן שלג

## 4 בונוס: משוב (+1)

על-סמל השבועות הראשונים של הסטטוס: אילו שינויים תרצו לראות באופן ההוראה על-מנת להבין את החומר יותר בקלות?

### 5 העשרה/הכנה: מידול שפהenganerative (0)

לא ציון, לא בדיקה. נדון זהה בשיעור. חומר עזר אפשרי: פרק 3 בסטנפורד.

מידול שפה הוא הבעיה של חיזוי הסתברות עבור טקסט נתון:

$$P(w_1, w_2, \dots, w_N)$$

ואשר לרוב נהוג לפרק לחיזוי מילה-מילה:

$$= P(w_1|\text{START}) \cdot P(w_2|w_1) \cdot \dots \cdot P(w_k|w_1, \dots, w_{k-1}) \cdot \dots \cdot P(w_N|w_1, \dots, w_{N-1}),$$

כאשר START הוא תבנית מיוחדת שלא מופיעה בטקסט ומסמנת תחילת מחרוזת.  
מודל שפה מסווג  $n$ -gram הוא מודל שבו אנו מפשטים את הנחת התלות של כל מילה במילים הקודמות,  
ומוחשבים רק את  $1 - n$  המילים שקדמו לה:

$$P(w_k|w_{k-n+1}, \dots, w_{k-1}).$$

את פרמטרי הסתברות נחשב לפי שכיחיות בקורס אימון, כך שנספר את כל רצפי המילים באורך  $n$  ונאמוד  
הסתברות באמצעות יחס ההופעה של המילה ה- $k$  במקומות  $1 - n$ :

$$\hat{P}(w_k|w_{k-n+1}, \dots, w_{k-1}) = \frac{\text{COUNT}(\langle w_{k-n+1}, \dots, w_{k-1}, w_k \rangle)}{\sum_w \text{COUNT}(\langle w_{k-n+1}, \dots, w_{k-1}, w \rangle)}.$$

1. האם זה מודל גנרטיבי (generative model)? הסבירו בקצרה.
2. אומדי הסתברויות של רצפים באורך  $n$  בלבד לא יאפשרו לנו לכלול בחישוב הסתברות הכללת את תחילת הטקסט. הסבירו בקצרה למה, והציגו אוסף נוסף של פרמטרים שנדרש לשומר כך שנוכל לכלול גם אותן. (יש יותר מושגשה אפשרית לנונה אחת.)
3. בהתאם לתשובתכם בסעיף הקודם,இיזה חלק יחסית של הפרמטרים מהמודל תופסים הפרמטרים מהאוסף החדש? בטאו כפונקציה של הגודלים הבאים או חלק מהם בלבד:  $n$ , גודל קורפוס האימון  $M$ , מספר המסמכים בקורס האימון  $D$ , גודל אוצר המילים  $V$ .
4. מה הסכנה שעולה לנבע מפיגישת  $n$ -gram בטקסט המבחן שלא הופיע בקורס האימון? הציעו שיטה אפשרית לפתורן הקשי.

Q 22)

ה	ו	ו	נ	ו	ב	כ	נ	#	
8	→ 7	→ 6	→ 5	→ 4	→ 3	→ 2	→ 1	→ 0	#
8	→ 7	→ 6	→ 5	→ 4	→ 3	→ 2	→ 2	1	ע
7	→ 6	→ 5	→ 4	→ 3	4	3	3	2	ו
7	7	→ 6	→ 5	→ 4	→ 3	4	3.5	3	ב
8	→ 7	→ 7	→ 6	→ 5	4	4	5	4.5	ו
9	→ 8	7.5	6.5	5.5	5	6	5.5	5	ת
10	→ 9	→ 8	7	6.5	6	7	6.5	6	ש
9	→ 8	9	8	7.5	7	8	7.5	7	ג
8	9	10	9	8.5	8	9	8.5	8	ה

Q 3)

אנו נזכיר:

$$f_1 = \begin{cases} 1 & \text{הוותק}, \\ 0 & \text{הוותק}. \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{הוותק}, \\ 0 & \text{הוותק}. \end{cases}$$

$$f_3 = \begin{cases} 1 & \text{הוותק}, \\ 0 & \text{הוותק}. \end{cases}$$

$$\vec{f} \left( \begin{matrix} \text{הוותק} \\ \text{הוותק} \\ \text{הוותק} \\ \text{הוותק} \\ \text{הוותק} \end{matrix} \right) = \langle 3, 2, 1, 1, 3.9 \rangle$$

$$\vec{\theta} = \langle 2.5, -5, -1.2, 0.5, 2.0, 0.7, 0.1 \rangle$$

$$\hat{y} = \sigma(7.5 - 10 - 1.2 + 0.5 + 7.8 + 0.1) \\ = \sigma(4.7) \approx 0.99$$

$$f_5 = \text{הוותק}$$

f<sub>1</sub> - Contains a Verb in past tense

OR

Contains Pronoun in "Second body"

$f_2$  - Contains Relation "o"

$f_3$  - # of words minus 5

$f_4$  - Contains "פָּתָח", "פָּנָס", "כְּנָס", "בְּלָשׁוֹן", "קַלְבָּה", "לְבָזָבָן"

$$1) \quad f\left(\begin{matrix} 150 & \text{ד} \\ \text{পৰ্যন্ত} & \text{জুলাই} \\ 20 & \text{১০/১৫/১৪} \end{matrix}\right) = \langle 1, 0, 1, 1 \rangle \quad y_1 = 1$$

$$f\left(\begin{smallmatrix} \gamma & \rho & \delta \\ \rho & \gamma & \delta \\ \delta & \delta & \gamma \end{smallmatrix}\right) = \langle 1, 0, 3, 0 \rangle \quad y_2 = 0$$

$$f \left( \begin{smallmatrix} r^k & l_3 \\ r'n & n \end{smallmatrix} \right) = \langle 0, 1, 0, 1 \rangle \quad y_3 = 1$$

$$2) : \theta^{(0)} = \langle 0, 0, 0, 0, 0 \rangle$$

$$h^{(1)} = 0.8$$

$$y^{(1)} = \delta(\Theta^{(0)} \vec{f} + b) = \delta(0+0+0+0+0) = 0.5$$

$$\hat{y} - y_1 = -0.5,$$

$$\nabla L_{CE}(\theta^{(0)}) = (-0.5, 0, -0.5, -0.5, -0.5)$$

$$\Theta^{(1)} = \langle 0.4, 0, 0.4, 0.4, 0.4 \rangle$$

$$\vec{\theta}^{t+1} = \vec{\theta}^t - \eta \nabla L_{CE}(\theta^{(t)})$$

$$\hat{y} = \sigma(\theta^{(1)} \vec{f})$$

$$\hat{y} = \sigma(0.4 + 0 + 1.2 + 0 + 0.4)$$

$$\hat{y} =$$

$$\hat{y} - y_2 = 0.5$$

$$\nabla L_{CE}(\theta^{(1)}) = 0.4$$

















