עיבוד שפה טבעית שבוע 6:
חזרה

פתרון הבוחן

הודעות

- יצאה **גרסה חדשה** של SLP - שימו לב שהישנה עוד נגישה מהאתר שלהם וכדאי לפנות אליה כדי לדייק בפרקים שלמדנו עד כה

5 העשרה/הכנה: מידול שפה באנגראמים (0)

ללא ציון, ללא בדיקה. נדון בזה בשיעור. חומר עזר אפשרי: פרק 3 בסטנפורד.

מידול שפה הוא הבעיה של חיזוי ההסתברות עבור טקסט נתון:

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

ואשר לרוב נהוג לפרק לחיזוי מילה־מילה:

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

כאשר START הוא תבנית מיוחדת שלא מופיעה בטקסט ומסמנת תחילת מחרוזת.

מודל שפה מסוג n-gram הוא מודל שבו אנו מפשטים את הנחת התלות של כל מילה במילים הקודמות, ומחשיבים רק את $n-1$ המילים שקדמו לה:

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

את פרמטרי ההסתברות נחשב לפי שכיחויות בקורפוס אימון, כך שנספור את כל רצפי המילים באורך n ונאמוד הסתברות באמצעות יחס ההופעה של המילה ה־ k במקום ה־ n:

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

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

1. האם זהו מודל גנרטיבי (generative model) ? הסבירו בקצרה.
2. אומדני הסתברויות של רצפים באורך **בדיוק** n בלבד לא יאפשרו לנו לכלול בחישוב ההסתברות הכוללת את תחילת הטקסט. הסבירו בקצרה למה, והציעו אוסף נוסף של פרמטרים שנידרש לשמור כך שנוכל לכלול גם אותו. (יש יותר מתשובה אפשרית נכונה אחת.)
3. בהתאם לתשובתכם בסעיף הקודם, איזה חלק יחסי של הפרמטרים מהמודל תופסים הפרמטרים מהאוסף החדש? בטאו כפונקציה של הגדלים הבאים או חלק מהם בלבד: n, גודל קורפוס האימון M, מספר המסמכים בקובץ האימון D, גודל אוצר המילים V.
4. מה הסכנה שעלולה לנבוע מפגישת n-gram בטקסט המבחן שלא הופיע בקורפוס האימון? הציעו שיטה אפשרית לפתרון הקושי.

4 העשרה / הכנה: הגדרת בעיות כתיוג רצפים (0)

ללא ציון, ללא בדיקה. נדון בזה בשיעורי החזרה.

תופעת **חילוף הקוד** (code switching), או לעתים **עירוב הקוד** (code mixing), מתייחסת למצב בו דובר.ת רב־לשוני.ת משתמש.ת במילים משתי שפות שונות בתוך אותו מבע (משפט). להלן כמה דוגמאות מעברית־אנגלית (העברית מתועתקת לטובת נוחות הכתב. האות x משמשת לציון ההגה ח, או כ לא־דגושה, ו־ š היא ההגה ש):

alon lakax et ha shoe off.

we limroxed the chocolate on the cake.

ma hi nag'a be?

ha witch kova šeli muxan.

1. הצביעו על קושי אפשרי בניתוח תחבירי של משפט מעורב־קוד, בליווי דוגמא שתמחיש חוסר יכולת להוציא עץ תלויות תקין. ניתן להשתמש בכל זוג שפות שהוא, לרבות בדוגמאות לעיל, ובלבד שהדוגמא תהיה מובנת ודקדוקית במידה סבירה עבור דובריהן. **שימו לב:** הכוונה בשאלה זו אינה ליכולת לזהות את המילים עצמן. הניחו מנתח תחבירי שיודע מה שפת המקור של כל מילה ומה חלק הדיבר שלה.

גישה יותר צנועה להתמודדות עם עירוב קוד היא להתייחס למבע מעורב כאל בעיית תיוג רצפים. בהינתן מבע מעורב, נתייג כל מילה לשפת המקור שלה.

2. האם יש צורך בסכימת תיוג מסוג BIO עבור הבעיה כפי שהוגדרה? נמקו.

3. הציעו סכימה בעלת **ארבעה** תגים לפחות להתמודדות עם תופעות שונות בדוגמאות לעיל, והדגימו את הצורך בכולם.

4. הניחו שלרשותנו דאטא עירוב קוד עברי־אנגלי הנתון בתעתיק כמו בדוגמאות לעיל ומתויג לפי סכימת התיוג שהגדרתם בסעיף הקודם. אנו מפתחים מערכת תיוג רצפים מבוססת־פיצ'רים עבור הבעיה (למשל, CRF). הציעו ארבע משפחות פיצ'רים (feature templates) למערכת התיוג, והסבירו בקצרה כיצד כל משפחה יכולה לעזור לפעולת החיזוי.

שאלות דוגמה - רב-ברירה (4-5 נק' כ"א)

א. המשחק וורדל (Wordle) מבקש מהמתחרה בו לנחש מילה אנגלית בת חמש אותיות תוך 6 נסיונות, כאשר בכל נסיון המשחק מסמן עבורה בצבעים שונים את: (א) האותיות שמצאה את מיקומן הנכון, (ב) אותיות שנמצאות במילה אבל לא במיקומן הנכון, ו-(ג) אותיות שאינן במילה כלל. המשחק מתעלם מסוג האות (case-insensitive). איזו מפקודות ה-shell הבאות **לא** תסייע לנו בבואנו להתחרות במשחק? הניחו שב-unix יש קובץ ובו כל המילים בשפה האנגלית (בכל האורכים), מילה בכל שורה.

A. פקודת diff
B. פקודת tr
C. פקודת grep
D. פקודת len (שסופרת אורך של מחרוזת) בתוך סקריפט awk

ב. איזו מפונקציות האקטיבציה להלן אינה גזירה בכל מקום?

A. ReLU
B. סיגמויד (sigmoid)
C. טנגנס היפרבולי (tanh)
D. העדר אקטיבציה

ד. באיזו מערכת אין הלימה בין המשימה שהיא מאומנת לבצע לבין המשימה שמעניינת אותנו?

A. מודל מרקובי סמוי לתיוג חלקי דיבר (HMM part-of-speech tagger)
B. רשת נוירונים נשנית לסיווג (RNN classifier)
C. מנתח תחביר רכיבי (constituency parser) ל-CFG
D. מנתח תלויות במעברים (transition-based dependency parser)

ה. אימנו מודל עבור משימה כלשהי, וניכר כי הוא סובל מהתאמת-יתר (over-fitting). איזו מהשיטות הבאות **לא** אמורה לעזור לו אם נוסיף אותה וננסה שוב?

A. רגולריזציית L2 (regularization)
B. דרופאאוט (dropout)
C. עצירה מוקדמת (early stopping)
D. הגדלת מימד חבוי (hidden dimension)