# עיבוד שפות טבעיות - תרגיל בית ראשון

## פרטי הגשה


**מגישים**

| | |
|---|---|
| אורי אריאל ניפומניאשצ'י | 318261468 |
| ישי גרוניך | 208989186 |


**מערכת הפעלה**

Windows 7

## חלוקת טקסט למשפטים

### חוקים כלליים לזיהוי סיום משפט

באופן כללי, תו יזוהה כסוף משפט אם אחד התנאים הבאים מתקיימים:

- זהו התו האחרון בטקסט
- זהו התו האחרון בפסקה (כלומר אחריו יש שורה חדשה)
- התו הוא אחד מהתווים (?, !,.)
  - כאשר ישנו אחד התווים (?, !,.) ואחריו התו ("), אז המשפט נגמר בתו ("). לדוגמה: הסטודנט קרא בשמחה: "כמעט סיימתי את תרגיל הבית!"
    - פירוט נוסף על ניקוד בקרבת מרכאות, באתר האקדמיה ללשון העברית

### מקרים יוצאי דופן עבור התו (.)

ישנן שימושים של התו (.) שאינם לצורך סיום משפט, אלא על מנת לסמן מילה מקוצרת. למשל:

- היום השקעתי 6 שעות בפתירת ש.ב. במקצועות שונים.
- ג. יפית: נוכחות טורדנית ומעצבנת. (מקור: Ynet)

מאחר שכל הקיצורים הללו הם קיצורים בעלי אות בודדת לפני הנקודה, זיהינו אותם על ידי בדיקת התו שנמצא שני-תווים לפני התו (.). כלומר, במקרה של המשפט הראשון, הנקודה הראשונה זוהתה כחלק מהמשפט (ולא כסוף משפט) מאחר שהתו הלפני-אחרון היה רווח ( ), ואילו הנקודה השנייה זוהתה כחלק מהמשפט (ולא כסוף משפט) מאחר ששהתו הלפני-אחרון היה נקודה נוספת (.). אם, במקום זאת, היתה אות נוספת, אז היינו מתייחסים לנקודה בתור סוף-משפט.

### התו הבעייתי (;)

לתו (;) ישנם מספר שימושים, אשר מפורטים באתר האקדמיה ללשון העברית.
אחד השימושים מציין סוף משפט, למשל:

היום אכלתי ארוחת צהריים בבית סבתי; היה מאוד טעים!

שימוש אחר אינו מציין סוף משפט, למשל:

היא הכינה לי מיני מאכלים: נעל מטובלת במלח גס, עם עלי תרד ממולאים; קבבים מעץ, מבושלים במיץ פטל משובח; תפוחי אדמה אדומים בציפוי דק של שוקולד מנטה.

לכן, החלטנו להניח שהשימוש בתו (;) הוא שימוש שאינו מפריד בין משפטים, כלומר כל שימוש של התו (;) הדומה לדוגמה הראשונה הנ"ל, לא יסווג בצורה נכונה.

## חלוקת משפטים לטוקנים Tokenization

### חוקים כלליים לזיהוי סיום token

באופן כללי, תו יזוהה כתו המסיים טוקן אם הוא עונה על אחת הדרישות:

- הוא התו האחרון במשפט
- אחריו יש רווח ( )
- התו הוא סימן המהווה טוקן בפני עצמו, או שמיד אחריו יש סימן המהווה טוקן בפני עצמו
  - "סימן המהווה טוקן בפני עצמו", הוא כל סימן שאינו רווח, ספרה, או אות - וזאת מלבד המקרים יוצאי הדופן המפורטים מטה.

### בעיות עם הסימן (-)

לסימן (-) ישנם שימושים שונים:

- המאחד בין שתי מילים או שני חלקי-מילה. בשימוש זה, אין רווח שני החלקים המחוברים. דוגמאות:
  - היום ביקרתי במסעדה, והאוכל היה תת-רמה.
  - היום ביקר בישראל המלך ג'ורג' ה-3.
- המפריד בין שני חלקים של משפט. בשימוש זה, יש רווח לפחות בצד אחד של הסימן. דוגמה:
  - לא הבנתי את מה שהמרצה הסביר - הייתי מאוד עייף ומחשבותיי גלשו לדברים אחרים.

במקרה הראשון נתייחס לשתי המילים (או יותר משתיים) בתור טוקן יחיד. במקרה השני נתייחס לסימן (-) בתור טוקן בפני עצמו. נבדיל בין שני המקרים, כאמור, על ידי הרווחים לפני ואחרי הסימן (-).

לסימן (-) ישנן צורות שונות:
ישנם תווים רבים באוסף התווים UTF-8 שכולם דומים לתו (-). בעברית, ישנם שני תווים בעלי משמעות שונה - האחד נקרא "מקף" (־) והשני נקרא "קו-מפריד" (-). ניתן לקרוא על כך עוד באתר האקדמיה ללשון העברית.
בפועל, השימוש הנפוץ ביותר, במיוחד באינטרנט ובפרט באתר Ynet, הוא שימוש בתו (-) בכל מקרה. לכן, הנחנו בקוד שלנו, שאין בכתבה תווים הדומים לתו (-), מלבד התו (-) עצמו.

## טיפול בתו (")

לסימן (") יש שתי מטרות שונות:

- סימון תחילת ציטוט או סוף ציטוט. במקרה זה התו (") הוא טוקן בפני עצמו.
- סימון ראשי תיבות (ר"ת). במקרה זה, התו (") אינו טוקן בפני עצמו, אלא חלק מהטוקן הכולל של ראשי התיבות.

נבדיל בין שני המקרים הללו על ידי העובדה שבמקרה הראשון, התו הבא אחרי (") אינו אות (או אינו קיים, כלומר זהו התו האחרון במשפט), ואילו במקרה השני התו העוקב חייב להיות אות.

## התו הבעייתי (')

לתו (') יש שימושים שונים:

1. סימון אות שנהגית בצורה שונה, למשל במילה "צ'יטה".
2. סימון של מילה בודדת מקוצרת: אונ', עמ', מס', וכו'.
3. התייחסות לאות בעברית, מבלי לכתוב את שם האות. למשל: א' = אלף.
4. סימון התחלה או סוף של מילה או ביטוי, שמשמעותן מושאלת. למשל: "לעיתים אני מרגיש שאני 'מדשדש במקום'.

בשלושת המקרים הראשונים, התו (') אינו טוקן בפני עצמו. במקרה האחרון, המופעים של התו (') הם אחד טוקנים בפני עצמם.

הבעייתיות היא, שלעיתים קשה להבדין בין שימושים (2,3) לבין שימוש (4). לדוגמה:

אבא שלי הכין לי היום לבית הספר 'סנדויץ' לוהט'.

לכן, נניח שבכל פעם שהתו (') נמצא בין שתי אותיות הוא חלק מהטוקן הכולל (כלומר נסווג בצורה נכונה את שימוש מספר 1) ונניח שבכל פעם שהתו (') אינו נמצא בין שתי אותיות, הוא טוקן בפני עצמו (כלומר נסווג בצורה נכונה את שימוש 4, אבל לא את שימושים 2+3).

## טיפול בתו (.)

התו (.) יכול להיות מיקומים שונים במשפט:

- בסוף המשפט, כמו ברוב המשפטים.
- בתוך קבוצה בסוף משפט, כמו במשפט "לעולם לא נדע..."
- באמצע משפט, כמו "קיבלנו היום ש.ב. במתמטיקה דיסקרטית."
- בתוך מספר עשרוני או תאריך.

עבור המקרה הראשון, נאמר שהתו (.) הוא טוקן בפני עצמו.

עבור המקרה השני, נאמר שקבוצת התווים (...) היא טוקן.

עבור המקרה השלישי, נאמר שהתו (.) הוא חלק מהטוקן "ש.ב.".

עבור המקרה האחרון, התו (.) הוא חלק מהמספר/מהתאריך.

כלומר, כדי לבדוק עבור תו מסוים (.) האם הוא טוקן בפני עצמו, נבדוק אם משני צדדיו יש ספרות, ואם לא, נבדוק האם כל התווים ממנו ועד סוף המשפט הם התו (.).

### טיפול בתו (,)

בדרך כלל התו (,) מסמן הפרדה בין חלקי משפט, כלומר מהווה טוקן בפני עצמו, אך לפעמים הוא נמצא בתוך מספר שמספר ספרותיו גדול, למשל 2,500. לכן על מנת לבדוק האם הוא טוקן בפני עצמו נבדוק אם משני צדדיו ישנן ספרות.

### טיפול בתו (/)

לתו (/) שני שימושים נפוצים:

- הפרדה בין שתי מילים או שני ביטויים שיש ביניהם יחס של "או". למשל במשפט: "לא ניתן לעלות למטוס עם כלי נשק ו/או חפצים העלולים לשמש ככלי נשק".
- שימוש כמפריד בין המספרים השונים של תאריך: 29/3/2016.

במקרה הראשון נתייחס לתו (/) כטוקן בפני עצמו, ואילו במקרה השני נתייחס אליו כחלק מהטוקן המייצג את התאריך. נזהה את ההבדל בין שני המקרים, באמצעות העובדה שבמקרה השני ישנה ספרה לפני התו '/' וגם אחרי התו '/'.

### טיפול בתו (:)

לתו (:) יש שני שימושים נפוצים, האחד בתוך מחרוזת המייצגת שעה 20:34, והשנייה כמפרידה בין שני חלקי משפט. כלומר בדומה לתו (/), התו (:) ייחשב כטוקן בפני עצמו בכל המקרים בהם הוא לא נמצא בין שתי ספרות.