תרגיל בית 1 – קורפוסים

פורטנוי ליאור
אפראימוב אורן

חלוקה למשפטים

האלגוריתם הנבחר לחלוקת משפטים, אשר מתבסס שהקלט הינו קלט תקין אשר עבר הגהה ואין בו ביטויים או תווים לא חוקיים, הינו:

- במידה ואלגוריתם רואה את הביטוי הבא ". " (נקודה ולאחריו רווח), הוא מחלק את המידע לשורה חדשה
- במידה ורואה את הביטוי הבא "? " או "! ", הוא מחלק את המידע לשורה חדשה.
- במידה ואין את ". " או נקודה במשפט (בדרך כלל בכותרות), האלגוריתם מפריד לשורות גם את השורה אחרונה (או היחידה) באותו אלמנט.

בגלל שזה מאמר ב-Ynet, האלגוריתם מתבסס על החלוקה של המאמר לפי האלמנטים בעץ ה-html ולכן לא יעבוד במקומות אחרים ששם זה אחרת, או שצורת החלוקה אחרת.

קשיים

- טיפול בתו ';' – סימן זה יכול להגיע במקום נקודה כאשר ישנו צורך להפריד בין שני משפטים (או כמה) שיש ביניהם קשר ענייני.
  סימן זה יכול להגיע במקום הסימן ',' כאשר הוא מגיע על מנת לפרט ביטוי במהלך המשפט.
- טיפול בתו ':' – סימן זה יכול להופיע לפני דיבור ישיר או ציטוט, במקרים אחרים יכול להגיע על מנת לפרט שורה של דברים (באמצעות נקודות)
- האלגוריתם נתקל בקושי רב של הוצאת כותרות משנה מגוף הכתבה, וזאת בגלל העובדה שאין אלמנט (של עץ האלמנטים – html) יחיד אשר מגדיר כותרות אלו. לכן, בגלל שיש אלמנטים רבים, אשר גם שייכים ל"תוכן זבל" (פרסומות, מטא-דאטה וכד'), האלגוריתם נתקל בקושי רב בהפרדה וזאת בגלל שאנו לא מבצעים ניתוח לשוני ואינינו יכולים לדעת אם יש קשר בין המשפטים וכך להיפטר מ"תוכן זבל".

האלגוריתם לוקח בחשבון, שיכול להופיע התו '\r' או '\n', ולכן הוא משמיט אותם ושם את הסימון '\r\n' (על מנת לתמוך בכל מערכת הפעלה).
בקשיים אשר פורטו כאן, האלגוריתם מתעלם מתווים אלו ולא עושה שום דבר היות ויש כפל משמעויות.

<u>הערה:</u> במידה ואין רווח אחרי הסימן "." האלגוריתם שלנו לא יידע מתי מסיים המשפט היות ויש מילים, כמו ש.ב, שנקודה חלק מהם ולכן לא נוכל לנחש מתי מסתיים המשפט. לכן, עבור טקסט כזה, האלגוריתם שלנו לא יידע מתי לחלק באופן תקין את הטקסט למשפטים.

טוקניזציה

תהליך הטוקוניזציה מומש על ידי הוספת הסימן רווח לפני ואחרי כל תו שיש צורך להפריד, **<u>פרט</u>** למקרים הבאים:

- כאשר התו '-' מופיע בין שתי אותיות או בין אות למספר, או בין 2 מספרים.
  לדוגמא: בית-הספר, ה-ספר, ב-3, 100-300 וכד'.
- כאשר התו מרכאות (") מופיע בין 2 אותיות.
  לדוגמא: צה"ל, מנכ"ל וכד'
- כאשר התו גרש (') מופיע בין 2 אותיות.
  לדוגמא: ג'יפ, ג'ירפה או שם של אדם
- כאשר התו '.' מופיע בין 2 ספרות.
  לדוגמא: 3.14, 02.02.15.
- כאשר התו ',' מופיע בין 2 ספרות.
  לדוגמא: 1,000, 1,000,000.
- כאשר התו '/' מופיע בין 2 ספרות.
  לדוגמא: 02/02/15.
  כאשר התו ':' מופיע בין 2 ספרות.
  לדוגמא: 13:33.

מכיוון שקיימות בשפה העברית (אם לא הרבה), מילים אשר מסתיימות ב-', אנו נתקלים בקושי בין הפרדה של המילה סנדוויץ' לבין ציטוט, ולכן בשני המקרים אנו נפריד ברווח למרות שבמקרה של הסנדוויץ' לא היינו צריכים (מילים אלו נדירות ומופיעות לא בתדירות גבוהה).

בשאר המקרים, אנו מפרידים על ידי רווח בין מילה לתו המיוחד.