**ממ"ן 14**

**חלק א'**

שאלה 1

מדדי Precision ו-Recall עבור השוואת סוגריים הניבו תוצאה 0, מדד FScore להשוואה זו לא הצליח להניב תוצאה (כנראה עקב חלוקה ב-0).

שאלה 2

למשפטים עד 40 מילים הדיוק בתיוג הוא 18.53%.

**חלק ב'**

שאלה 1

שאלה 2

**אלגוריתם:**

1. לכל צומת parent שמהווה אב ליותר מ-2 בנים:
   1. צור תת-עץ sub שבניו שני הבנים הימניים ביותר של parent והאב תווית לפי "כלל התוויות".
   2. לכל בן child של parent מימין לשמאל חוץ משני בנים אלו והבן השמאלי ביותר:
      1. צור תת-עץ tmp ש-child בן שמאלי בו ו-sub בן ימני בו והאב תווית לפי "כלל התוויות".
      2. הצב את tmp ב-sub.
   3. מחק מ-parent את כל הבנים חוץ מהבן השמאלי ביותר.
   4. הוסף את sub כבן ימני של parent.

**"כלל התוויות" עבור התווית בשורה 1.1:**

מחרוזת בה יש משמאל לימין את תווית האב והתוויות של כל הבנים מופרדות במקף ('-'), למעט שני הבנים הימניים ביותר שמשמאלם במקום מקף יש כוכבית ('\*')

**"כלל התוויות" עבור התווית בשורה 1.2.1 בבן child:**

מחרוזת בה יש משמאל לימין את תווית האב והתוויות של כל הבנים מופרדות במקף ('-'), משמאל ל-child שבמקום מקף יש כוכבית ('\*')

**מוטיבציה:**

כל אב רב-בנים ובניו מומרים לתת-עץ ייחודי שגם ניתן להחזרה באופן חד-חד ערכי ועל. התוויות הייחודיות מונעות טעויות ספירה.

**חלק ג'**

פונקציית המטרה:

אלגוריתם (מבוסס על האלגוריתם בספר הקורס עם תוספת לחוקים אונריים):

1. בכל תא CYK[i-1][i] הצב את ההסתברות הלקסיקלית של מילה מספר i במשפט.
2. לכל j מ-1 עד אורך המשפט:
   1. לכל i מ-j-2 בירידה עד 0:
      1. לכל s מ-i+1 עד j-1:
         1. לכל תווית מתאימה X:
            1. הצב בתא CYK[i][j][X] את הערך המתקבל מפונקציית המטרה.
            2. הצב בתא BP[i][j][X] את (Y,Z)
            3. הצב בתא S[i][j][X] את s
            4. עבור רקורסיבית על תוויות של חוקים אונריים שבצד ימין יש את X והוסף לשלושת הטבלאות את התוויות שמשמאל

סיבוכיות עבור מספר משפטים n, אורך משפט s, מספר תגים t (בסדרי גודל):

* זמן ריצה: כי יש שלוש רמות של לולאות שעוברות על כל משפט ושתי רמות שעוברות (במקרה הגרוע) על כל תג ואז רמה נוספת לבדיקת חוקים אונריים.
* זיכרון: כי צריך טבלה שבכל תא יכול להיות כל אחד מהתגים וצריך עץ לכל משפט שמכיל מידע כאורך המשפט.

**במקום חלק ד' וחלק ה'**

בתהליך הפיתוח, בדקנו את הפונקציות השונות על משפטים בודדים לטובת בדיקת הקוד.

בשלב ההרצה הסופי נתקלנו בבעיות יעילות קשות – מאות שניות לשורה.

התעסקנו רבות בייעול הקוד, אך לצערנו הרב לא הגענו למצב שאנחנו מצליחים לרוץ על כלל העצים בזמן סביר.

תהליך הפיתוח מתועד כאן:

<https://github.com/aszarf/maman14_NLP>