מבוא לעיבוד שפה טבעית

מטלה 12 - פתרון

רומן ציפרון, ת.ז. 306591645

שאלה 1

1. בהנחה כי "מופעי יוניגרם של סגמנטים" מתייחס למספר הסגמנטים בקובץ (כולל תווים מיוחדים כגון נקודות, פסיקים, סימני קריאה), יש 127884 סגמנטים בקובץ האימון ו-11282 סגמנטים בקובץ הגולד.
2. בהנחה כי "סוגי יוניגרם של סגמנטים" מתייחס למספר הסגמנטים השונים בקובץ, יש 15986 סמגנטים שונים בקובץ האימון ו-3171 סגמנטים שונים בקובץ הגולד.
3. בהנחה כי "מופעי סגמנט-תג" מתייחס למספר התיוגים של סגמנטים (שאינם תווים מיוחדים – משמע שורות 30 – 1 ללא שורה 31 בטבלה 3 במאמר *treebank*), יש 113044 מופעי סגמנט-תג בקובץ האימון ו-9897 מופעי סגמנט-תג בקובץ הגולד.
4. בהנחה כי "סוגי סגמנט-תג" מתייחס למספר התיוגים השונים בקובץ של סגמנטים שאינם תווים מיוחדים, יש 26 תיוגים שונים בקובץ האימון ו-25 תיוגים שונים בקובץ הגולד.
5. מדד העמימות עבור קובץ האימון: 1.13493056424
   מדד העמימות עבור קובץ הגולד: 1.07978555661
6.