תרגיל בית 1 – קורפוסים

חיים שאללשוילי 200832780

צחי כפיר 200681476

חלוקה למשפטים

האלגוריתם הנבחר לחלוקת משפטים:

כל משפט יסתיים במחרוזת המתאימה לביטוי הרגולרי הבא : [("")]*[?!.][?!.] בתנאי והמחרוזת אינה נקודה המוקפת בספרות משני צדיה או נקודה המוקפת באותיות משני צדיה. ביטוי זה יפריד כל משפט המסתיים ב:נקודה , סימן קריאה וסימן שאלה, משפט המסתיים ברצף הבנוי מאותם תווים, ומשפט המסתיים ברצף המסתיים באחד מהתווים הללו ואחריהם מירכאות כפולות.

הסיבה לכך היא שבהבאת דיבור ישיר בין מירכאות מצטרפות שתי מערכות פיסוק: הפיסוק של הטקסט המצוטט והפיסוק של הכותב. ראוי למנוע גיבוב של סימני פיסוק רצופים, ועל כן יש לוותר על סימן פיסוק השייך לאחת המערכות . כדי לבצע זאת סימן הפיסוק יבוא לפני המירכאות למשל:

הוא פרש זרועותיו וקרא: "ריבונו של עולם"!

קשיים:

- טיפול בתו ";" – סימן זה יכול להגיע במקום נקודה כאשר ישנו צורך להפריד בין שני משפטים שיש ביניהם קשר ענייני אמיץ ובמרום פסיק כאשר יש לפרט עניינים ין האיברים השונים בחלקי המחרוזת. בשל כפל המשמעות הוחלט לא להפריד משפטים על פי תו זה בשלב זה כאשר אנו לא משתמשים בידע מורפולוגי וסמנטי על הטקסט.
- טיפול בתו ":"– במקרים מסוימים סימן זה מופיע לפני ציטוט או דיבור ישיר ובמקרה זה אין צורך לרדת שורה. במקרים אחרים שבהם התו נקודתיים מגיע לפני פירוט של נקודות נרצה להפריד למשפט חדש. בשל כפל המשמעות הוחלט לא להפריד משפטים על פי תו זה בשלב זה כאשר אנו לא משתמשים בידע מורפולוגי וסמנטי על הטקסט.
  החלטה זו לא פוגמת ברוב הטקסטים שכן טקסטים שבהם קיימים רשימות של נקודות קיימים כבר סימני ירידת שורה שאנו לא מוחקים באלגוריתם שלנו. לכן בכל מקרה תתבצע חלוקה למשפט חדש.
- כן האלגוריתם לקח בחשבון כי בטקסט המקורי קיימת הפרדה בסיסית של רווח בין משפט למשפט. שכן במקרה ולא, קיימת האפשרות כי נגנוב תווים מהמשפט הבא לדוגמא מירכאות כפולות.
- סימנים שונים של ירידת שורה עבור כתבנים שונים

טוקניזציה

תהליך הטוקניזציה ממוש ע"י הוספה של רווח (' ') לפני ואחרי כל תו שאיננו אות, פרט למיקרים הבאים:

- כאשר התו '-' מופיע בין שתי אותיות. לדוגמא: בצירופים תל-אביב, בית-ספר אין צורך להוסיף רווחים, אך במידה ומופיע הטקסט הבא: "מקרה ראשון-" נרצה להפריד את "ראשון" ואת '-'
- כאשר התו '"' מופיע בין שתי אותיות. לדוגמא: בראשי התיבות צה"ל, חז"ל אין צורך להוסיף רווחים, אך במידה ומופיע ציטוט, דני אמר לי: "מסור לי את הכדור" , נרצה להפריד בין ה- '"' "מסור" "כדור" '"'
- כאשר התו '" מופיע בין שתי אותיות. לדוגמא: במילים ג'יפ, ג'ירפה, אין צורך להוסיף רווחים, אך במידה ומופיע הטקטס הבא: קראתי את הספר 'מלחמה ושלום', נרצה להפריד בין '" "ושלום".

  ישנו קושי להבדיל בין שלום' מהמשפט הנ"ל לבין במילה סנדויץ'. בשני המיקרים התו '" לא מופיע בין 2 אותיות אך פעם אחת צריך להוסיף רווח ופעם אחת לא. לבסוף החלטנו כי רק כאשר התו '" מופיע בין 2 אותיות לא להוסיף רווחים, כייון שבעברית מילה לא יכולה להתחיל בתו זה, ובמיקרים נדירים מאוד מילה מסתיימת בתו, וגם לאחר הוספת הרווח לא נוצר בלבול עם מילה אחרת.
- כאשר התו '.' מופיע בין שתי ספרות. במקרה של נקודה עשרונית $\pi = 3.141592653589$, אין צורך להוסיף רווחים, אך במקרה של נקודה בסוף משפט נרצה להוסיף רווחים.
  כמו כן במקרה והתו מופיעה בהקשר של תאריך 14.3.1989
- כאשר התו ',' מופיע בין שתי ספרות. במקרה של ייצוג מספרי של מספר גדול 1,000,000 , אין צורך להוסיף רווחים, אך במקרה של פיסוק במשפט נרצה להוסיף רווח.
- כאשר התו '/' מופיע בין שתי ספרות. במקרה תאריך 14/3/89 , אין צורך להוסיף רווחים, במקרה ומופיע בין שתי אותיות נרצה להוסיף רווח. לדוגמא כאשר מופיע הטקטס: זכר/נקבה נרצה להפריד בין "זכר" '/' "נקבה".

המילים נפוצות ביותר בקורפוס הבדיקה

1	,	52587
2	.	43495
3	של	18151
4	"	16331
5	-	13148
6	*	12395
7	את	11159
8	(	10513
9	)	10493
10	על	9709
11	:	5089
12	הוא	5036
13	עם	3553
14	ב	3459
15	גם	2837
16	לא	2627
17	בשנת	2593
18	או	2446
19	'	2422
20	היא	2283
21	היה	2229
22	ידי	2100
23	לאחר	2068
24	בין	2000
25	זה	1955
26	כי	1650
27	אך	1502
28	יותר	1450
29	כל	1429
30	אשר	1386
31	עד	1344
32	ה	1305
33	זו	1239
34	הם	1218
35	ישראל	1199
36	הייתה	1139
37	בית	1136
38	כמו	1100
39	חיצוניים	1063
40	קישורים	1047
41	ביותר	1041
42	;	997
43	אחד	992
44	שם	982
45	הראשון	967
46	בו	964
47	מספר	940
48	אותו	924
49	היו	874
50	יש	867

אחוז המילים ששכיחותן מעל 50 הוא 1.8%
אחוז המילים ששכיחותן מתחת ל-10 הוא 89.7%
אחוז המילים ששכיחותן בדיוק 1 הוא 54%

ישנם 197981 מופעים של 10 המילים הראשונות, 21.4% מכלל הטקסט
ישנם 269433 מופעים של 50 המילים הראשונות, 29.1% מכלל הטקסט
ישנם 301497 מופעים של 100 המילים הראשונות, 32.6% מכלל הטקסט

הנתונים מתיישבים עם חוק Zipf כפי שניתן לראות בגרף: