עיבוד שפות טבעיות - תרגיל 1 - קורפוסים

שלב 1 - טיפול בטקסט

.1

בסעיף זה יצרנו מחלקה בשם PROTOCOL שמכילה:

שדה שיכיל את שם הקובץ,שדה שיכיל את מספר הכנסת ,אחד שיכיל סוג הפרוטוקול, שדה שיכיל נתיב לקובץ ועוד שדה שיכיל רשימה לשמירת שמות המדברים והמשפטים.

.2

בסעיף זה מימשנו שתי פונקציות כל אחד מטפלת באחד מסוגי הפרוטוקולים שיש לנו ומתייחסת בהתאם.

זיהינו את שמות המדברים ע"י חיפוש ":" או "<<" או "<" בסופו של טקסט ואז בדקנו לפי סטייל ולפי קו תחתון ובולד.

אם השם מתחיל בכל אחד מהמילים הבאות:

כמו פרופ׳ , היו"ר (שנכתבת עם סוג " שונה),ראש הממשלה,המשנה לראש הממשלה

,סגן שר במשרד ראש הממשלה ,

נשיא הפרלמנט האורופי (אין עוד צירוף שמתחיל במילת נשיא אז זה נחשב כמקרה חריג) ועוד כמה כאלה

כל ביטוי שמתחיל ב "השר" צירפנו לאוסף גם כן מכיוון שלא מצאנו חוקיות משותפת (יש יותר משני מקרים וכל מקרה יש לו מעט הופעות)

במקרה ומתחילה ב "שר" אכן מומש קוד בהתאם מלבד כמה מקרים חריגים .

אז הוא מחק אותן בהתאם

אח"כ טיפלנו במקרה בו השמות מתחילות ב "תשובת" ומחקנו.

-בדקנו אם מה שנשאר מתחיל ב "סגן","שר","סגנית","שרת","מזכיר הכנסת","מזכירת הכנסת" שרוב השמות אכן מכילות אחת מהן ואז מחקנו והתחלנו לחפש איך לנקות שאר השם לפי חוקיות שזיהינו : אם המילה מתחילה באות "ה" במקרה ויש עוד האם ההיא מסתיימת ב "," ויש אחריה עוד מילה שמתחילה ב "ה" ואז ישנה מילה שמתחילה ב "וה" ומוחקים כל מה שהיה עד כה , במקרה ואין "," בודקים במקום אם המילה שאחרי ב "וה" ומוחקים.

.b

הבעיות שיכולות להיות בשימוש בשמות כפי שהופיעו בפרוטוקולים לפני שנמחקו:

-אי סדר

-ישנם מלא מקרים שהשם רשום עם התפקיד המלא שלו ואז בהמשך הוא יהיה ללא התפקיד ולכן ברגע שאני מנסים למצוא את המשך דיבוריו לא נמצא כי התפקיד אינו נמצא

הבעיות שיכולות להיות בשימוש בשמות אחרי הניקוי:

-יכול לקרות ששם יאבד את חלק משמו הפרטי מכיוון שתחילת השם שלו הינה תפקיד (במקרה של מר למשל אם נתקלנו בשם שכולל מר אז הוא מוחק את זה והשם יהיה לא שלם)

- איבוד משמעות (במקרה והניקוי לא היה תקין)

-חפיפה בין שני אנשים עם תפקידים שונים שיש להם אותו שם

.3

קבענו לזהות את גבולות בין משפטים ע"י זיהוי "." , "?" או "!" שמעידים על סיומו של משפט.

.4

לאחר שקיבלנו לכל מדבר את המשפטים ששיכים אליו ,

עברנו על כל משפט מחקנו ממנו כל הופעה של " - -" (זיהינו שני סוגים של קווים – נכללו בבדיקה שלנו כמובן) ,גם כן מחקנו על הופעה של אותיות באנגלית (מימשנו פונקציה הבודקת אותיות באנגלית בהתאם),גם כן בוצע ניקוי של סימנים כמו <,>.

.5

"!","?","." מעידים על סיומו של משפט ולכן החלטנו שיהיו טיקן בנפרד.

"," מחקנו אותם פרט למקרים שהם נמצאים במספרים.

( ) נחשבו כטוקן גם כן.

"-" השארנו כדי לשמור על משמעות הצירוף (למשל ראש-הממשלה אם נפריד כל אחת לבד זה ייאבד משמעות הצירוף).

.6

בוצעה בדיקה לגבי אורך הטוקן ונכללו רק אלה שאורכם 4 לפחות.

.7

עברנו על כל 100 הקבצים ,ושמרנו בהתאם לכל טוקן שאורכו לפחות 4 את המידע הבא (שהם השדות במחלקת פרוטול שכבר הזכרנו בסעיף הראשון)

עבור כל קובץ שמרנו ע"י חילוץ את שם הקובץ ,מספר הכנסת וסוג הפרוטוקול.

ואז שמרנו שם הדובר וכל משפט שייך אליו ע"י כך שפונים למיקום המתאים ששומר כל אחד מהם.

שלב 2 – מימוש חוקZIPF

.2

המשמעות של הגרף היא:

הגרף מתאר את התפלגות שכיחות המילים בטקסט

ניתן לראות בגרף שכמות קטנה של מילים הן נפוצות מאוד

המילים שנמצאות באמצע הגרף שומרות על היחס הלינארי לפי חוק זה.

.3

כן הגרף תואם את הציפיות שלנו.

לפי חוקZIPF

Let fn be the frequency of the n-th word on the list (whose rank is n) Then there exists a constant k such that fn ×n = k

נוכל לראות שהחוק מתקיים לפי הגרף הסקלה לוג לוג

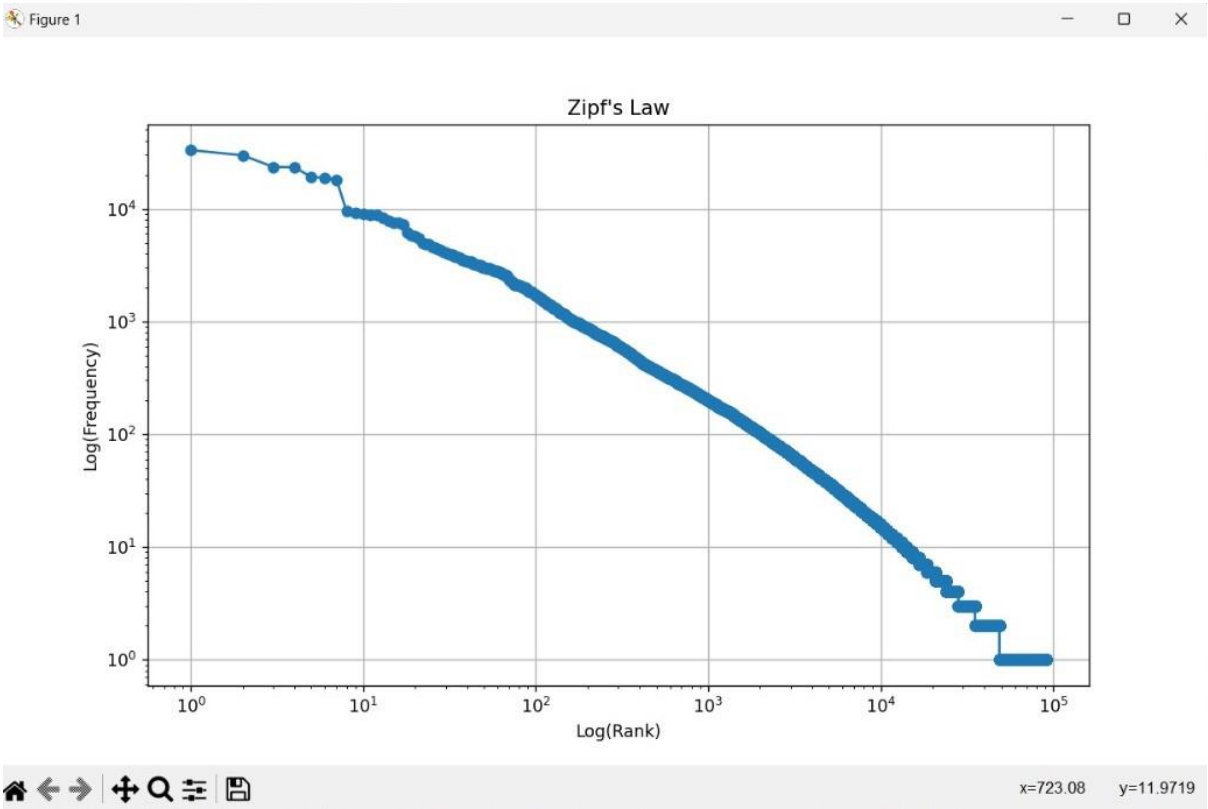
הגרף נראה לינארי לכן מתקיים החוק.

.4

הגרף היה נשאר אותו דבר ,מכיוון שהחוק הינו אונברסלי והוא אכן תקף לא משנה כמה גודל הקורפוס שלוקחים.

אכן אינו משפיע.

.5



.6

Most common tokens:[את,לא,אני,של,זה,על,הכנסת,חבר,גם,הוא]

כן תואמות, מכיוון שמילים אלו שקיבלנו הינם מילות קשר אשר משתמשים בהם לרוב בכל משפט בשפה עברית בנוסף לכך יש מילת כנסת שאכן כל הפרוטוקולים הינם פרוטוקולי כנסת ואכן מובן מאליו שמילה זו תחזור על עצמה כמה פעמים בכל פרוטוקול.

Least common tokens:
[נתלית,מונחה,הטלה,מתאוששת,דילגנו,משויפים,נצלול,טעימה,שנאתם, הסובייקטיבי]

אכן גם אלה תואמות מכיוון שאלו מילים שאינן שימושיות בשפה העברית באופן כללי.