טל חכם סורקיס 315095398
עומר יצחקי 312230022

דו"ח תרגיל 1 – קורפוסים

שלב 1 – טיפול בטקסט

2. שליפת טקסט בעל תוכן.

התבקשנו לשלוף מכל פרוטוקול את הטקסט הרלוונטי. טקסט רלוונטי הוגדר לנו כשמות הדוברים והטקסט אשר שייך לכל דובר.
שמנו לב כי בכל פרוטוקול, המלל הראשוני אינו רלוונטי לנו, לכן רצינו למצוא בפרוטוקול את ההופעה הראשונה של דובר כלשהוא, כאשר מצאנו הופעה ראשונה של דובר- מכאן עד לסוף הפרוטוקול כל הטקסט רלוונטי לנו. אמנם בגוף הפרוטוקול עצמו ישנן כותרות ומידע שאינו רלוונטי, אך מכיוון שניתנה לנו יד חופשית לתוכן זה, החלטנו לצרף כותרות אלו למלל של אותו דובר אחרון בפרוטוקול השייך לו. בחירה זאת נבעה מכך שכתיבת הקוד הייתה פשוטה יותר, כי ידענו להפריד בין מלל של דוברים שונים בכך שידענו לזהות כאשר הופיע דובר חדש בפרוטוקול.

כיצד זיהינו דובר בוועדה/מליאה ?

ראשית עברנו על פרוטוקולים שונים ושמנו לב כי בכל פעם כאשר הופיע דובר, הוא היה מסומן בקו תחתון ונקודתיים. לכן רצינו למצוא היכן מופיע תכונה זאת של קן תחתון (underline). תחילה חשבנו שזה אמור להיות תחת "תכונה" בודדת, מכיוון שמדובר בקו תחתון. אך לאחר שמצאנו את אותה "תכונה", שמנו לב כי אנו מפספסים שמות רבים של דוברים. כאשר עשינו debug גילינו כי אותה "תכונה" מופיעה במקומות מסוימים כ-NONE . לכן הבנו כי ככל הנראה הקו תחתון נמצא ב"תכונה" שונה. לבסוף גילינו למעשה 3 מקומות שונים שמאפיינים את ההופעה של קו תחתון. כלומר חיפשנו ב-3 מאפיינים שונים בקוד האם המאפיין הינו TRUE. במידה ואחד משלושת המאפיינים הללו הינו TRUE זה העיד לנו על קו תחתון. כעת רצינו לבדוק הופעה של נקודתיים עם השילוב של הופעה של קו תחתון.
עשינו זאת אך גם כעת שמנו לב כי לא כיסינו את כלל שמות הדוברים, וזאת מכיוון שבחלק מן הפרוטוקולים ישנם דוברים בודדים שאכן מופיעים עם קו תחתון אך ללא נקודתיים. מכיוון שהבחנו כי זהו פחות מאחוז בודד מכלל הדוברים, החלטנו לא להתייחס לדוברים אלו. בדקנו כי הנקודתיים נמצאים בסוף המשפט (סוף שם הדובר) או שהנקודתיים מוצגים בחבילה Document עם אחד הסימנים: > < .

טל חכם סורקיס 315095398
עומר יצחקי 312230022

שמנו לב כי ישנם משפטים שעונים על שתי הדרישות הללו, אך לא מדובר בשמות של דוברים. לכן יצרנו רשימה של חלקי המשפטים הללו בכדי שלא נתייחס אליהם כשמות של דוברים.

כיצד ביצענו את ניקיון השמות של הדוברים?

שמנו לב כי כאשר שמות הדוברים לא היו נקיים, לכל דובר יש תפקיד מסוים, אשר היינו צריכים לנקות. בין אם מדובר בתפקיד של הדובר או תואר כזה או אחר. בפרוטוקולים שמות המפלגה, תפקיד והתואר של האדם הופיעו לפני שם הדובר (או בסופו אך זה יהיה מלווה עם סוגריים), לכן חשבנו בהתחלה לקחת פשוט את 2 המילים האחרונות בכל שם דובר. אך מהר מאוד גילינו שזהו פתרון שאינו מכסה שמות של דוברים רבים, וזאת מכיוון ששמות רבים של דוברים אינם מוגדרים ב-2 מילים, וכי יש שמות רבים שהם 3 מילים, 4 מילים ואף יותר. לכן הבנו שאנו צריכים לחשוב על פתרון יצירתי יותר.

לשם כך בנינו ביטוי רגולרי (regex) אשר מטרתו הינה להסיר את המילים שאינם מגדירות את שמו של האדם. הבחנו בחוקים הבאים ולפיכך בנינו את הregex:

<> בתחילת וסוף השם, תשובת, תפקיד, תיק, סוגריים, נקודותיים.

ה- regex מטפל בניקיון השמות ונותן לנו את השמות הפרטיים ושמות המשפחה של הדוברים.

בעיות שהופיעו בשימוש בשמות כפי שהופיעו בפרוטוקולים לפני ואחרי הניקיון:

לפני:

לפני ניקוי השמות שמו של דובר יכול להופיע בשלל ורציות שונות (לדוגמא: השר לבטיחות בדרכים ישראל כץ, שר האוצר ישראל כץ, תשובת השר ישראל כץ וכו') ובכך למרות שמדובר בדובר בודד לפני ניקוי השמות היינו מקבלים שמדובר בדוברים שונים. אנו חושבים שזה יכול להוות בעיה אם היינו רוצים לעשות סטטיסטיקה על הנתונים לפי שם הדובר.

אחרי:

לאחר ניקוי השמות הופיעה לנו בעיה כאשר שמות מסוימים נמחקו בחלקם: מרב מיכאלי (לפני הניקוי), ב מיכאלי (לאחר הניקוי) שרונה פלדמן (לפני הניקוי), ונה פלדמן (לאחר הניקוי). גילינו זאת בכך שעברנו על קובץ השמות שלנו. מכיוון שעבדנו עם ביטוי רגולרי

שבחלקו מחק מילים כמו מר/שר. ולכן השמות לאחר הניקוי הופיעו לא נכון, תיקנו את בעיה זו בכך שדרשנו שלאחר תפקיד יהיה חייב לבוא רווח.

3. חלוקה למשפטים.

כיצד זיהינו גבולות בין משפטים בתוך הטקסט:

ראשית נאמר כי כאשר עבדנו על תרגיל בית זה השתמשנו במידע מהאתר של האקדמיה ללשון העברית. וזאת מכיוון שכפי שנאמר בהרצאות, בנייה מוצלחת יותר של קורפוס נובעת מידע של אותה השפה. לאחר בדיקה באתר של האקדמיה ללשון העברית מצאנו לנכון כי התווים אשר מעידים על סיום משפט הינם: נקודה (.), סימן קריאה (!), סימן שאלה (?) ונקודה פסיק (;). בנוסף בפרוטוקולים ישנם משפטים רבים אשר מסתיימים ללא סימן פיסוק, אלא עם שורה חדשה (enter) לכן החלטנו לחלק את הפרוטוקולים למשפטים כאשר אנו מזהים את אחד מסימני הפיסוק הנ"ל או שורה חדשה. עשינו זאת ע"י בנייה של רשימה, אשר מזהה את אחד הסימנים הללו. אך יש סימן פיסוק אחד שהינו זקוק לטיפול מיוחד וזהו הנקודה. כאשר נקודה מופיעה כחלק מתאריך או כסימון של מספר כלשהו, או כאשר הנקודה בדיוק לפני סוף של ציטוט (מלווה עם מרכאות) לא חילקנו למשפט חדש.

4. ניקיון המשפטים.

כיצד זיהינו וניקינו משפטים שאינם תקינים:

עברנו על כל המשפטים בפרוטוקולים, ובכל משפט עברנו אות אות ובדקנו ראשית שלא מופיעה אות באנגלית, שקיימת לפחות אות אחת בעברית, וכי אין את הסימן מקף שחוזר על עצמו פעמיים רצוף במהלך המשפט. במידה ושלושת התנאים הנ"ל מתקיימים המשפט הינו תקין והוספנו אותו לקורפוס, אחרת לא התייחסנו למשפט ודילגנו עליו.

מקרה קצה: היו משפטים שנראים תקינים במסמכי הוורד אך כאשר ניתחנו את הפרוטוקולים ב- python קיבלנו כי חלק מן המשפטים קיים הסימון <<>> וזאת על פי הספרייה document. בגלל שקריאת המשפטים הללו כפי שהם מוצגים בוורד תקינים החלטנו להשאיר אותם כחלק מהקורפוס שלנו.

טל חכם סורקיס 315095398
עומר יצחקי 312230022

5. טוקניזציה.

מכיוון שלא התבקשנו להתייחס למורפמות, העבודה לחלק לטוקנים הייתה פחות מורכבת. למעט מקרים חריגים שנפרט בהמשך, חילקנו את המשפט לטוקנים כאשר הופיע רווח או סימן פיסוק או אנטר. לדוגמה המשפט: טל ועומר אוכלים 5 במבות בשבוע.

יחולק כך:

טל

ועומר

אוכלים

5

במבות

בשבוע

.

הערה: בקובץ, הטוקנים חולקו ברווחים כמבוקש, אך לצורך נראות בדו"ח נכתב באופן הנ"ל.

מקרים חריגים שבהם לא חילקנו לטוקנים:

כאשר זיהינו כי בטקסט יש תאריך (שמלווה עם סימן פיסוק: נקודה או קו נטוי) החשבנו אותו כטוקן אחד ולא הפרדנו אותו למספר טוקנים, אותו הדבר גם לגבי תיאור של השעה (שמלווה עם סימן פיסוק: נקודתיים) וגם מספרים כגון 100,000 ואחוזים (80%) וכסף ($80).

את סימן הפיסוק " החלטנו לא בהכרח להפריד כטוקן מכיוון שמילים כמו: יו"ר, סכו"ם. לא היינו רוצים שיופרדו לטוקנים: יו ר סכו ם

אך לעומת זאת כאשר סימן הפיסוק " מעיד על תחילה או סיומת של ציטוט כן רצינו שיפריד כטוקן. לכן בקוד בדקנו האם סימן הפיסוק " מעיד על מרכאות או על ראשי תיבות. במידה וזה חלק מראשי תיבות לא הפרדנו לטוקן, ואחרת כן הפרדנו לטוקן.

טל חכם סורקיס 315095398
עומר יצחקי 312230022

שלב 2 – מימוש חוק zipf

2. משמעות הגרף.

גרף zipf שיצרנו מציג את שכיחויות המילים מהפרוטוקולים שקיבלנו כתלות לינארית בדרגה שלהן. לאחר שהשתמשנו בקנה מידה לוגריתמי הן בציר X והן בציר Y וזאת כדי לקבל גרף שהוא לינארי.
כך שהמילה השכיחה ביותר נמצאת בתחילת הגרף והמילים הנדירות נמצאות בסוף.

3. האם הגרף תואם את הציפיות שלנו ?

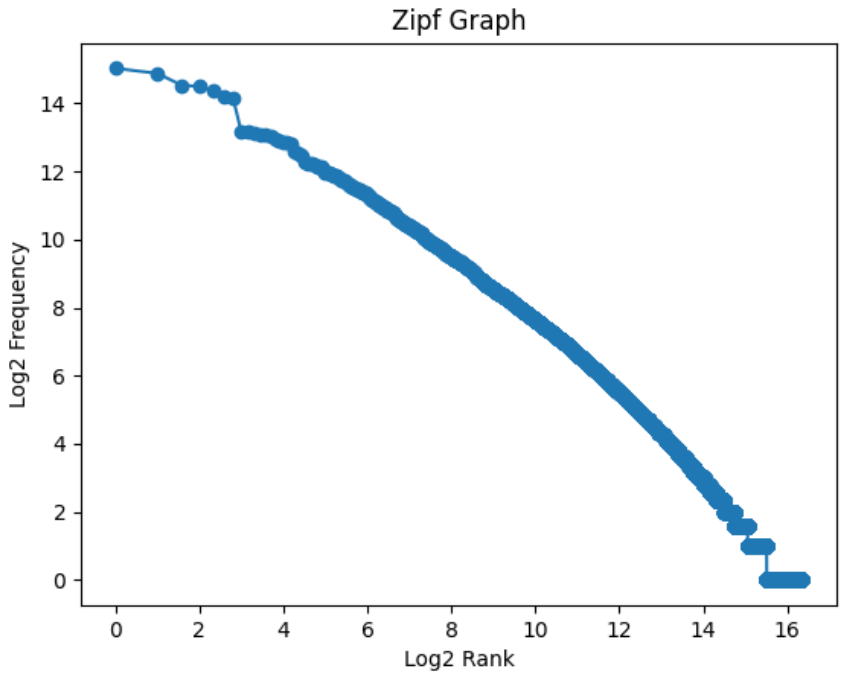
כן. כפי שכתבתנו בסעיף 2, ניתן לראות בגרף שלנו את התלות הלינארית. קיבלנו אכן גרף שתואם לציפיות כפי שהוצג לנו בגרף של Moby Dick מההרצאה.

4. מה היה קורה לגרף אם היינו מקטינים/מגדילים את גודל הקורפוס?

כפי שנאמר בהרצאה, בין אם הקורפוס הינו קטן יותר או גדול יותר לא יהיה הבדל משמעותי לנראות הגרף.

5.

6.



טל חכם סורקיס 315095398
עומר יצחקי 312230022

עשר המילים עם התדירות הכי גבוהה בסדר יורד:

את, לא, של, אני, על, זה, הכנסת, חבר, גם, הוא

עשר המילים עם התדירות הכי נמוכה:

נתלית, מונחה, הטלה, מתאוששת, דילגנו, משויפים, נצלול, טעימה, שנתאם, הסובייקטיבי

הערה: ישנן מילים רבות שיש להן הופעה בודדת, אלו 10 מתוכן.

האם המילים תואמות את הציפיות שלנו ?

לגבי המילים עם התדירות הכי נמוכה לא היה לנו ציפיות לאיזה מילים דווקא כן יופיעו, לעומת זאת לא ציפינו לקבל מילים שמשתמשים באופן תדיר בשפה העברית כגון: מילות קישור, מילות גוף. ואכן לא קיבלנו מילים כאלו.

לגבי המילים עם התדירות הכי גבוהה, אכן ציפינו לקבל מילות קישור וגוף(ראשון/שני/שלישי) וכך קיבלנו כי חלקן מופיעות בעשר מילים עם התדירות הכי גבוהה.
יחד עם זאת חשבנו כי המילה "אני" תופיע בשלוש מילים הראשונות, מכיוון שחברי הכנסת במדינת ישראל אוהבים לדבר על עצמם. הופתענו (חלקית) לגלות כי "אני" הגיעה למקום הרביעי. בנוסף ציפינו לקבל את המילים "חבר" ו"הכנסת" זאת משום שבכל זאת מדובר בפרוטוקולים של הכנסת.

הערות נוספות שהתבקשנו לפרט על פלט שגוי:

אלו מקרי קיצון של תופעות לשוניות אשר הקשו עלינו בפיתוח קוד גנרי שתופס את כל המקרים.

1. באחד מן הפרוטוקולים הופיע המשפט הבא ב"מבצע שלמה"
אצלנו הפלט לאחר חלוקה לטוקנים הינו ב"מבצע שלמה "
היינו רוצים שהפלט יהיה ב " מבצע שלמה "
זוהי דוגמה לכך שהפלט אינו כפי שהיינו רוצים וזאת מכיוון שראשית לא התבקשנו להתייחס

טל חכם סורקיס 315095398
עומר יצחקי 312230022

למורפיזם (ב-) אך יתרה מכך זה קורה בגלל שהתייחסנו למרכאות כציטוט רק כאשר אין אות עברית גם מימין וגם משמאל (כפי שהסברנו בנ"ל לגבי מרכאות כראשי תיבות).

2. הפרדת שורות שלא היינו רוצים כדוגמת א.ד גורדון

אצלנו הפלט לאחר חלוקה לשורות הינו:

א.

ד גורדון

אך מכיוון שא.ד מציין את שמו הפרטי היינו רוצים לקבל:

א.ד גורדון

3. מלל בדוגמאות: 1. עומר 2. טל 3. אגוז

אצלנו הפלט לאחר חלוקה לשורות הינו :

1.

עומר 2.

טל 3.

אגוז

אך היינו רוצים שהחלוקה לשורות תהיה כך:

1. עומר
2. טל
3. אגוז