תרגיל 1 – קורפוסים

מבוא

בתרגיל זה נבנה קורפוס, כלומר, ניצור מאגר טקסטואלי נרחב איתו נעבוד במהלך הקורס. הטקסט איתו נעבוד הוא טקסט **בעברית,** הלקוח מפרוטוקולים של מליאות וועדות הכנסת. הטקסט נכתב ברובו על ידי חברי כנסת, שרים ואורחים בוועדות הכנסת. הטקסטים איתם נעבוד מורכבים הן משפה דבורה שנכתבה (ונערכה מעט) על ידי קלדנים/יות והן מנאומים כתובים מראש.

שלב 1 – טיפול בטקסט

לתרגיל מצורף קובץ zip המכיל 100 מסמכי *word* בפורמט docx. כל מסמך מהווה פרוטוקול אחד. עליכם ליצור ממסמכים אלו **קורפוס** באופן הבא (תוך פירוט על בחירותיכם בדו"ח שתכתבו):

1. <u>שליפת נתונים מתוך קבצי הפרוטוקולים:</u> כל שם של מסמך הוא בפורמט הבא XX_ptv_fileNumber.docx **או** XX_pt**m**_fileNumber.docx. כאשר:
   a. XX – הוא מספר הכנסת אליו הפרוטוקול שייך.
   b. ptm – מסמל שזהו פרוטוקול של מליאה.
   c. ptv – מסמל שזהו פרוטוקול של ועדה.

   עליכם לשלוף ולשמור לכל שם קובץ:
   א. את מספר הכנסת אליו הוא שייך כInteger.
   ב. אינדיקציה אם זה פרוטוקול של מליאה או של ועדה:
      a. עבור וועדה ערך השדה צריך להיות הstring: "committee"
      b. עבור מליאה ערך השדה צריך להיות הstring: "plenary"
2. <u>שליפת טקסט בעל תוכן:</u> כל פרוטוקול מתחיל לרוב בסימונים, כותרות, פירוט סדר היום, רשימת מוזמנים וכו'. מבחינתנו הטקסט הרלוונטי הוא זה השייך לדוברים בוועדה/מליאה. חישבו על דרך להבחין בין הטקסטים הרלוונטיים לשאר הטקסט ושילפו מתוך כל פרוטוקול את שמות הדוברים, כפי שהופיעו בפרוטוקול והטקסט השייך לכל דובר/ת. כיתבו בדו"ח על ההחלטות שלכם ודרך המימוש שבחרתם.
   a. שימו לב ששמות הדוברים יכולים להופיע עם תוספות כמו תפקיד, שם המפלגה וכו'. עשו ככל יכולתכם לנקות את התוספות ולהשאר רק עם שם הדובר/ת. פרטו בדו"ח איך ביצעתם את הנקיון הנ"ל.
   b. חישבו אילו בעיות יכולות להיות בשימוש בשמות <u>כפי שהופיעו בפרוטוקולים</u> לפני ואחרי נקיון השמות. ענו על כך בדו"ח.

   כדי לבצע את שלב זה, אתם יכולים להשתמש במחלקה *Document* מתוך הספריה *docx* על מנת לעבור על פסקאות ולקרוא את הטקסט מתוך המסמך.
   דוגמה <u>פשוטה</u> לשימוש בספריה:

```
from docx import Document

document = Document(file_path)

for par in document.paragraphs:
    par_text = par.text
```

3. חלוקה למשפטים: לאחר ששלפתם לכל דובר את כל הטקסט השייך לו, עליכם לקבוע כיצד לזהות גבולות בין משפטים בתוך הטקסט, ולפרט על קביעתכם בדו"ח.
   - **אין להשתמש בספריות חיצונית לחלוקת המשפטים**
4. נקיון המשפטים: חלק מהמשפטים בקורפוס יכולים להיות לא תקניים. למשל משפטים באנגלית או משפטים שמכילים רק תווים שאינם אותיות. כמו כן, ישנם גם משפטים שנחתכו באמצע (לא שלמים) ולרוב מסומנים ע"י תווים דוגמת "- - -". נסו לזהות ולנקות משפטים כאלו ולהשאר רק עם משפטים תקינים בעברית. דווחו איך התמודדתם עם משימה זו בדו"ח.
5. טוקניזציה: עליכם לקבוע כיצד לחלק משפטים לטוקנים ולממש זאת על הטקסט (התמודדות עם סימני פיסוק, ראשי תיבות ועוד). יש לפרט על קביעתכם בדו"ח.
   - למעט מקרים חריגים, סימני פיסוק יהיו טוקנים נפרדים. חשבו על המקרים החריגים שיש להתייחס אליהם שונה ופרטו על כך בדו"ח.
   - אין צורך לבצע ניתוח מורפולוגי למילים, כלומר, **אין צורך להפריד מורפמות**, למשל ריבוי (כמו "ספרים"), אותיות חיבור ("וספר"), ה' הידיעה ("הספר") וכו'.
   - כל טוקן מופרד ברווח אחד.
   - **אין להשתמש בספריות חיצונית לטוקניזציה**
6. שליפת משפטים איתם ניתן לעבוד: בקורס נעסוק בשפות טבעיות, וכדי לחקור אותן נרצה להשתמש בקורפוסים המורכבים מצירופי מילים, ולא מילים בודדות. לכן, נכלול בקורפוס רק משפטים שבהם לפחות 4 טוקנים.
7. שמירת הנתונים כקובץ CSV: הפלט של התוכנה צריך להיות קובץ csv עם העמודות הבאות:
   a. protocol_name: שם הקובץ של הפרוטוקול (ראו סעיף 1.1).
   b. knesset_number: מספר הכנסת ממנה הפרוטוקול לקוח (ראו סעיף 1.1).
   c. protocol_type: האם הפרוטוקול הוא ועדה או מליאה (ראו סעיף 1.1).
   d. speaker_name: שם הדובר (ראו סעיף 1.2).
   e. sentence_text: משפט השייך לאותו דובר לאחר טוקנזיציה (ראו סעיפים 1.2-1.6).
   - קובץ הCSV יכיל את כל המשפטים שכללתם בקורפוס, מכל הפרוטוקולים יחדיו.
   - ניתן להשתמש לשם כך בספריית *pandas*.
   - **יש לכתוב את הפלט בקידוד utf-8.**

שלב 2 – מימוש חוק zipf

בשלב זה בידכם טקסט נקי ומופרד לפי טוקנים. נרצה לבדוק אם חוק Zipf מתקיים עבור הקורפוס שיצרתם. לשם כך:

1. ממשו פונקציה שמחשבת ומשרטטת גרף (plot) המציג את חוק Zipf על הטוקנים שבקורפוס שיצרתם.
   a. ציר הX מייצג את לוג הדרגה $\log(rank)$ של הטוקנים.
   b. ציר הY מייצג את לוג התדירות $\log(frequency)$ של הטוקנים.
   - התעלמו מטוקנים שאינם מילים (סימני פיסוק וכדו').
   - ניתן להשתמש בספרייה Matplotlib לייצור הגרף.
2. הסבירו מה המשמעות של הגרף.
3. האם הגרף תואם את הציפיות שלכם? הסבירו.
4. מה היה קורה לגרף אם היינו מקטינים את גודל הקורפוס? ומה אם היינו מגדילים?
5. צרפו תמונה של הplot שקיבלתם לדו"ח.
6. הדפיסו את רשימת 10 המילים עם התדירות הכי גבוהה שקיבלתם ואת 10 המילים עם התדירות הכי נמוכה. האם המילים תואמות את הציפיות שלכם? צרפו את רשימות המילים לדו"ח.

הערות כלליות

1. עבור סעיף 1.2 ייתכן ותצטרכו להתמודד עם כותרות שמופיעות באמצע הפרוטוקול. טקסטים אלו אינם שייכים לאף דובר. תוכלו לבחור איך אתם מתמודדים עם טקסטים אלו- אם למשל לצרף אותם כטקסט של הדובר האחרון, כטקסט של היו"ר או להתעלם מהם לחלוטין. כיתבו בדו"ח את בחירתכם והסבירו.
2. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות עבור כל שלב בתהליך. השתמשו ב Try-Except blocks לפי הצורך.
3. לשם נוחות, אני מציעה ליצור Class בשם Sentence המכיל את הפרטים ברמת המשפט, וClass נוסף בשם Protocol המכיל את הפרטים ברמת הפרוטוקול ובתוכו רשימה של איברים מסוג Sentence. זאת הצעה אך אתם <u>לא</u> מחוייבים לכך.
4. אתם יכולים לעבוד בכל סביבת עבודה שנוחה לכם, אך הפתרון ייבדק בסביבת windows ועליכם לדאוג שהוא ירוץ בהצלחה בסביבה זו.

אופן ההגשה

1. ההגשה היא בזוגות בלבד.
2. עליכם להגיש <u>קובץ zip</u> בשם <id1>_<id2>.zip (כאשר <id1>,<id2> הם מספרי תעודות הזהות של הסטודנט הראשון והשני בהתאמה), המכיל את הקבצים הבאים:

a. קובץ python בשם **processing_knesset_corpus.py** המכיל את כל הקוד הנדרש כדי לממש את שלב 1.
   i. - הקלט לקובץ יהיה נתיב לתיקיית מסמכי הdocx שקיבלתם ונתיב לשמירת הפלט.
      - הפלט צריך להיות קובץ הcsv כפי שמתואר בשלב 1.7.
   ii. על הקובץ לרוץ תחת הפקודה:

```
python processing_knesset_corpus.py <input_corpus_dir> <output_path>
```

b. קובץ python בשם **knesset_zipf_law.py** המכיל את כל הקוד הנדרש כדי לממש את שלב 2.
   i. - הקלט לקובץ הוא נתיב לקובץ הcsv שיצרתם בשלב 1 ונתיב לשמירת הפלט.
      - הפלט צריך להיות plot כפי שמתואר בשלב 2.1, השמור כקובץ תמונה.
   ii. על הקובץ לרוץ תחת הפקודה:

```
python knesset_zipf_law.py <input_csv_file> <output_path>
```

c. קובץ pdf בשם **<id1>_<id2>_hw1_report.pdf** ובו דו"ח המפרט על הקוד ועל ההחלטות שקיבלתם במהלך העבודה על התרגיל. כיוון שאנו מתעסקים בשפה טבעית וכל פרוטוקול שונה בתבניתו מפרוטוקולים אחרים, ייתכן ותגלו שהקוד שלכם הניב גם תוצאות לא צפויות שלא עומדות במה שקיוויתם. באופן כללי, עליכם לעשות מאמץ שהפלט יהיה טוב ככל הניתן, אך אין ציפייה לקבל פלט מושלם. בכל מקרה שהפלט שגוי, עליכם להתייחס לכך בדו"ח ולהסביר למה התופעה מאתגרת. שימו לב שהכוונה היא לא להסבר ברמת הקוד אלא ברמת <u>התופעה הלשונית או הטקסטואלית</u> שמקשה על פיתוח קוד גנרי שתופס את כל המקרים.
הדו"ח צריך להכיל גם מענה על השאלות בסעיפים השונים, וגם תמונה של הפלט ורשימת המילים מסעיף 2.
<u>אל תשכחו לציין בתחילת הדו"ח את שמותיכם ותעודות הזהות שלכם.</u>

d. קובץ הcsv שיצרתם בשלב 1 בשם **knesset_corpus.csv**

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד, כמו גם שימוש בכלי AI דוגמת chatGPT.

ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.

יש להגיש את התרגיל עד לתאריך 28.1.24 בשעה 23:59.

בהצלחה!