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 ו-ועדות 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 תיקייה שבה כ100 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 שמות קבצי הפרוטוקולים:</u> כל שם של מסמך הוא בפורמט הבא XX_ptv_fileNumber.docx **או** XX_ptm_fileNumber.docx. כאשר:
   a. XX – הוא מספר הכנסת אליו הפרוטוקול שייך (הכנסת הראשונה היתה ב-1948; הכנסת הנוכחית היא ה-25).
   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 כ-Integer.
   ב. אינדיקציה אם זה פרוטוקול של מליאה או של ועדה:
      a. עבור וועדה ערך השדה צריך להיות ה-string: "committee"
      b. עבור מליאה ערך השדה צריך להיות ה-string: "plenary"

2. <u>שליפת מספר פרוטוקול:</u> בתחילת כל פרוטוקול מצוין בד"כ מספר הפרוטוקול או מספר הישיבה. שלפו מספר זה מתוך הטקסט, עבדו אותו אם צריך ושמרו כ-Integer. אם לא קיים מספר כזה או לא הצלחתם לחלץ אותו שמרו כ- 1- .

3. <u>שליפת טקסט בעל תוכן:</u> כל פרוטוקול מתחיל לרוב בסימונים, כותרות, פירוט סדר היום, רשימת מוזמנים וכו'. מבחינתנו הטקסט הרלוונטי הוא זה שנאמר על ידי דוברים בוועדה/מליאה. חישבו על דרך להבחין בין הטקסטים הרלוונטיים לשאר הטקסט ושילפו מתוך כל פרוטוקול את שמות הדוברים, כפי שהופיעו בפרוטוקול, ואת הטקסט שנאמר על ידי כל 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 ולהישאר רק עם שם הדובר/ת. פרטו בדו"ח איך ביצעתם את הנקיון הנ"ל.

b. חישבו אילו בעיות יכולות להיות בשימוש בשמות כפי שהופיעו בפרוטוקולים 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 פשוטה לשימוש בספריה:

```
from docx import Document

document = Document(file_path)

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

הערה:

ייתכן ותצטרכו להתמודד עם כותרות או טקסטים אחרים שמופיעים באמצע הפרוטוקול. טקסטים אלו אינם משויכים לאף דובר. תוכלו לבחור איך אתם מתמודדים עם טקסטים אלו: למשל, לצרף אותם כטקסט של הדובר האחרון, כטקסט של היו"ר או להתעלם מהם לחלוטין. כתבו בדו"ח את בחירתכם והסבירו.

4. 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 חיצוניות לחלוקת המשפטים**

5. נקיון המשפטים: חלק מהמשפטים בקורפוס יכולים להיות לא תקניים. למשל, משפטים המכילים אנגלית, משפטים שמכילים רק תווים שאינם אותיות, משפטים שנחתכו באמצע (לא שלמים) ולרוב מסומנים ע"י תווים דוגמת "- - -" ועוד. נסו לזהות ולסנן משפטים לא תקינים כאלו וגם אחרים, ולהישאר רק עם משפטים תקינים בעברית, תוך התחשבות באילוצים. דווחו איך התמודדתם עם משימה זו בדו"ח.

6. 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 "ספר**ים**"), אותיות חיבור ("**ו**ספר"), ה' הידיעה ("**ה**ספר") וכו'.
- כל טוקן מופרד ברווח אחד.

- **<u>אין</u> להשתמש בספריות חיצוניות לטוקניזציה**

7. <u>שליפת משפטים איתם ניתן לעבוד</u>: בקורס נעסוק בשפות טבעיות, וכדי לחקור אותן נרצה להשתמש בקורפוסים המורכבים מצירופי מילים, ולא מילים בודדות. לכן, <u>נכלול בקורפוס רק משפטים שבהם לפחות 4 טוקנים</u>.

8. שמירת הנתונים כקובץ jsonl:

קובץ JSONL (JSON lines) הוא קובץ שבו כל שורה היא JSON תקני.

הפלט של התוכנה צריך להיות קובץ JSONL שבו כל שורה היא JSON של משפט אחד עם השדות הבאים:

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. protocol_number: מספר הפרוטוקול (ראו סעיף 1.2)
e. speaker_name: שם הדובר (ראו סעיף 1.3).
f. sentence_text: משפט השייך לאותו דובר לאחר טוקנזיציה (ראו סעיפים 1.3-1.7).

- קובץ הJSONL יכיל את כל המשפטים שכללתם בקורפוס, מכל הפרוטוקולים יחדיו.
- ניתן להשתמש לשם כך בספריות json, *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 $\log_2(rank)$ של הטוקנים.
   b. ציר הY מייצג את לוג התדירות $\log_2(frequency)$ 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. 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 לא מחוייבים לכך.
2. 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 התופעה הלשונית או הטקסטואלית שמקשה על פיתוח קוד גנרי שתופס את כל המקרים.
3. אתם יכולים לעבוד בכל סביבת עבודה שנוחה לכם, אך הפתרון ייבדק בסביבת windows עם python 3.9 ועליכם לדאוג שהוא ירוץ בהצלחה בסביבה זו.
4. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות עבור כל שלב בתהליך ולא לקרוס. השתמשו ב-Try Except blocks לפי הצורך.
5. שימו לב, בבדיקת תרגילי הבית בקורס ניתן משקל גדול מהניקוד הן על **הדו"ח**, ההסברים והידע שהפגנתם בחומר הנלמד והן על **הקוד**, אופן המימוש, יעילותו, קריאותו ועמידותו. בפרט, הרבה מהבדיקות הן אוטומטיות ולכן עליכם להקפיד על קוד תקין שרץ ללא שגיאות ועל עמידה **מדוייקת** בפלט הנדרש וביתר ההנחיות.
6. ניתן לשאול שאלות על התרגיל בפורום המיועד במודל. למעט מקרים אישיים מיוחדים, אין לשלוח שאלות הקשורות לתרגיל הבית במייל.
7. על אחריותכם לעקוב אחר הודעות הקורס במודל (בלוח הודעות ובפורום) ולהיות מעודכנים במידה ויהיו שינויים בהנחיות.

ספריות מותרות לשימוש

אתם יכולים להשתמש רק בספריות החיצוניות שהוזכרו במפורש בתרגיל:
pandas, docx, matplotlib
ובכל ספריה **סטנדרטית** של python.
אתם יכולים לחפש שם של ספריה ב https://docs.python.org/3/library/index.html על מנת לבדוק אם זו ספריה סטנדרטית. לא יהיה מענה על שאלות לגבי שימוש בספריות ספציפיות.

- למען הסר ספק, json היא ספרייה סטנדרטית של python.
- מומלץ להשתמש עבור כל פרוייקט בסביבה וירטואלית virtual environment חדשה משלו על מנת להיות בטוחים שאתם משתמשים רק בספריות מותרות ולמנוע קונפליקטים עם ספריות קודמות שהתקנתם בעבר. ראו מצגת על כך במודל.

אופן ההגשה

1. ההגשה היא בזוגות בלבד.
2. עליכם להגיש קובץ zip 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 ה-jsonl כפי שמתואר בשלב 1.7.
      ii. על הקובץ לרוץ תחת הפקודה (ללא הסימונים <>):

```
python processing_knesset_corpus.py <path/to/input_corpus_dir> <path/to/output_file_name.jsonl>
```

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

```
python knesset_zipf_law.py <path/to/input_file_name.jsonl> <path/to/output_file_name.png>
```

   c. קובץ pdf בשם **<id1>_<id2>_hw1_report.pdf** ובו דו"ח המפרט על הקוד ועל ההחלטות שקיבלתם במהלך העבודה על התרגיל. הדו"ח צריך להכיל גם מענה על השאלות בסעיפים השונים, וגם תמונה של הפלט ורשימת המילים מסעיף 2. ניתן להגיש את הדו"ח בעברית או באנגלית לנוחיותכם.
      אל תשכחו לציין בתחילת הדו"ח את שמותיכם **בעברית** ותעודות הזהות שלכם.
   d. קובץ ה-jsonl שיצרתם בשלב 1 בשם **knesset_corpus.jsonl**

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

ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.

יש להגיש את התרגיל עד לתאריך 23.5.24 בשעה 23:59.

בהצלחה!