

תרגיל 1 – קורפוסים – דוח מסכם

שלב 1 - טיפול בטקסט

מבוא

המטרה בתרגיל הייתה לבנות קורפוס בעברית מトーר כ80 פרוטוקולים של הכנסת (קבץ וורד בפורמט docx). המערכת קוראת כל מסמך, מחלצת מטא-דאטה שם הקובץ, מזהה דוברים ואת הנאומים שלהם, מחלקת למשפטים, מבצעת ניקוי וטוקניזציה בסיסית ללא ספריות חיצונית, ושומרת פלט לקובץ מסווג JSON (שורה = משפט) בקידוד UTF8. הדגש היה דיקב בבחירה הטקסטים הרלוונטיים והימנעות מסינון יתר, תוך שמיירה על עמידה מלאה בדרישות של התרגיל.

מטרות המטרה

- **חילוץ מטא-דאטה שם הקובץ:** מספר הכנסת (int) וסוג פרוטוקול (plenary/committee).
- **חילוץ מספר פרוטוקול:** זיהוי והחזרה כתו או 1 - אם לא נמצא.
- **חילוץ יוזר הישיבה:** איתור יוזר כשם פרטי+משפחה נקי מתארים.
- **חילוץ דוברים וtekstiyim:** שמיירה על טקסט רלוונטי בלבד והימנעות מכותרות/תוכן ענייניים.
- **חלוקת למשפטים וניקוי:** ללא ספריות חיצונית, עם סף מינימלי של לפחות 4 TOKENS.
- **טוקניזציה פשוטה:** פיסוק CTOKENS נפרדים (עם מספר מקרים חריגים), רוח יחיד בין TOKENS.
- **פלט JSON:** בפורמט קבוע, בקידוד UTF-8 ללא הדפסות למסך בכלל.

מתודולוגיה ויישום

(א) חילוץ מטא-דאטה שם הקובץ

- שם קובץ בפורמט: NN_ptv_ID.docx. או NN_ptm_ID.docx
 - ביטוי רגולרי: ax_\(\d{+}(_pt)([mv])\)_docx
- knesset_number:int = NN ○
protocol_type="plenary" = m ○
protocol_type="committee" = v ○
- אם הפורמט לא תואם - הקובץ מдолג באופן בטוח (לא קrise בעזרת try-except).

(ב) חילוץ מספר הפרוטוקול

- חיפוש במסמך כלו לפי תבניות נפוצות:
פרוטוקול מס, +פ\ מספר ישיבה; +פ\ ישיבה מס, +פ\ פרוטוקול; +פ\
- אם נמצא אך מבוצעת המرة ל *to* אחרת -.

(ג) זיהוי יי"ר הישיבה

- יי"ר מוגדר רק כשהנמצאת תגיית דובר אמיתי שמכילה ביטוי יי"ר (למשל הי"ר / ישב-ראש וכו').
- מנגן ניקוי שם (clean_name) מסיר תארים/תפקדים (למשל ח"כ/שר/ד"ר), סוגרים וסימני פיסוק מיוחדים, ודוחה מחרוזות שלא מייצגות שם (כותרות, ציטוטים, התחלת במיראות)
- אם לא זזה יי"ר מפורשות אך אנחנו שומרים מחרוזת ריקה.

(ד) זיהוי דוברים והפרדה מטקסטית מערכת

- שורת דובר חוקית: מסתiemת בנקודותים ונראית כשם לאחר ניקוי. כתורות/תוכן עניינים/הודעות מערכת מסוננים.
- הפקציה `looks_like_header` מזהה ומסננת: "תוכן העניינים", "הצעות לסדר-היום", "תאריכים/שעות/מקום (ירשלים, שעה...)", "מסמכים שהונחו...", "הודעת/סקירת...", "הצעת חוק...", "קריאה שנייה/שלישית" ועוד.
- לפני הופעת הדובר הראשון - לא מייחסים טקסט לאף אחד (גנעים מהדבקת פתיח לי"ר).
- לאחר מכן - רק טקסט שמופיע אחרי תגיית דובר מיוחס לדובר הנוכחי, שורות ללא דובר נכון אין משמעות.

(ה) חלוקה למשפטים

- פיצול על גבולות . ? ! ; (לא מתבצע פיצול שרואים נקודתיים ":")
- סינון רוחחים וצמצום מקטעים ריקים.

(ו) ניקוי משפטיים

- הסרה של רצפים לא תקינים: קווים תחתונים ---,תו נקודה תבליט •.
- יחס מינימלי של עברית במשפט (`hebrew_ratio` ≥ 0.25) כדי למנוע סינון יתר ולשמור כמות נתוניים.
- שמירה רק אם יש יותר מ 4 TOKENS.

(ז) טוקניזציה

- פיסוק נפרד לטוקנים (כללית).

- **חריגים נשמרים כמילה אחת:**
 - גרשים בין אותיות עבריות (למשל ח"כ/ו"ר) נשמרים כמילה אחת באמצעות הגנה זמנית.
 - מקף בתוך מילה עברית (למשל "יושב-ראש") נשמר.

(ח) פלט JSON

- כתיבה ל JSON-בקידוד utf-8 לא הדפסות למסך.
 - שימוש ב (..., ensure_ascii=False, json.dumps(..., ensure_ascii=False) לשמיירת עברית תקינה.
 - החזרת קודי יציאה ללא הדפסות.
 - 1 - שימוש לא תקין, 2 - תקינות קלט לא קיימת, 3 - שגיאת קריאה/סריקה, 4 - שגיאת כתיבה, 0 - הצלחה.

אתגרים ושיקולים

- **שמות שאינם דוברים (איכות הנתונים):** ביטויים כמו "مسקנות הוועדה", "מנהיגות הוועדה", "רשימת פרלמנטרית" הופיעו בתגיות של דוברים. פתרון: רשימת חריגים + סינון כותרות לפני שייר לדבורה.
 - **סינון לא אגרסיבי מדי:** הוגדר סף עברית מתון (0.25) ושמירה על ספרות/פיסוק, כדי לשמר על כמהות משמעותית של משפטים.
 - **דיק ביז"ר:** יז"ר נקבע רק מתחילה תגית דובר אמיתית עם ציון יז"ר, ולא מסרייקות כלליות, מנע טעויות כמו קטיעי פתיחים שנקלטו כשם יז"ר.
 - **טיפוסים נכונים:** protocol_number | knesset_number נשמרים כמספרים או כ-1 אם לא נמצא.
 - **יציבות והימנעות מקריסות:** סביר כל שלב יש try/except כדי למנוע קriseה של התהילה על קובץ בעייה'.

מבנה הפלט ודוגמה

שדות הפלט לכל משפט:

protocol_name, knesset_number, protocol_type, protocol_number,
protocol_chairmain, speaker_name, sentence_text

דוגמה (מරצת בדיקה):

{"protocol_name":"13_ptm_532058.docx","knesset_number":13,"protocol_type":"plenary","protocol_number":-1,"protocol_chairmain":null,"speaker_name":null,"sentence_text":null,"sentences":[]}]
חברי הכנסת, אנחנו פותחים את ישיבת הכנסת, יומ שישי, כ"ז אוקטובר 1992 בכסלו התשנ"ג, 22 בדצמבר

שלב 2 - שאלות

שאלה 1

אני חשב שההאמנת עניין של trade-off מצד אחד, פיצול המורפומות יכול לעזור למודל להבין טוב יותר את המבנה הדקדוקי - כי אז הוא רואה בבירור את ההבדל בין החלקים התפקידיים (כמו "ו" או "כש") לבין השורש של המילה עצמה. במקרה, זה יוצר איזשהו סטנדרט - אותו שורש תמיד יופיע באותה צורה, לא משנה אילו תחilibות מחוברות אליו.

מצד שני, יש בעיה ממשית: כשמפרקים מילה למורפומות, אפשר לאבד את המשמעות השלמה שלה. ניקח לדוגמה ביטוי כמו "וכשיבאו" - זה לא סתם רצף של מורפומות, זה ביטוי שלם שמתאר מצב ספציפי. גם מבחן טכנית, כל פיצול כזה פירושו יותר טוקנים, מה שמסרב אל העיבוד.

בהתחשב בדרישות של התרגיל, נראה לי שבבחירה לא לפצל היא די סבירה - היא חוסכת סיבוכים מיוחדים ושומרת על זרימה טبيعית יותר.

שאלה 2

נניח שהייתי מחליט לפצל את המילה **"לכשתצרכו"** היותי מחלק אותה באופן הבא:

- **ל** - מורפמת יחיד
- **כש** - מורפמת זמן
- **צרכו** - הפועל עצמוני

למה ככה? כי כל אחד מהחלקים האלה עובד אחרת בתוך המשפט. "ל" ו- "כש" זה בעצם כלים דקדוקיים שיכולים להתחבר למיללים שונים, ואילו "צרכו" זה הליבה - הפעולה שבאמת קורית. ככה המודל יכול **"ללמוד"** את המורפומות התפקידיות פעמי אחת ולהשתמש בהן בכל מיני הקשרים.

שאלה 3

להחלטה לשומר כל משפט בנפרד עם כל הפרטים שלו, יש כמה יתרונות ברורים:

קודם כל, זה נוח מאד לעבודה מדעית. רצים לבדוק מה מדובר במסויים אמר על נושא מסוים? פשוט מחפשים לפי השדות הרלוונטיים. גם מבחינות ניתוח - אפשר לזרוץ על המשפטים אחד אחד, לسؤال אותם, לבדוק טוון, מה שרצו.

אבל יש גם מחריר: הקשר בין המשפטים לא ממש נשמר. משפט שני עשוי להיות תשובה לשירה למשפט ראשון, אבל אם הם מאוחסנים כרשומות נפרדות, הקשר הזה לא אוטומטי. בנוסף, כל

משפט נושא אליו את כל המטא-דאטा (מספר פרוטוקול, דובר וכו'), מה שיוצר המון מידע כפול וחרדיות במערכת.

לעומת זאת:

- אם שומרים את כל הפרוטוקול ביחד, יש את ההקשר המלא אבל קשה יותר לבדוק משפטיים בודדים.
- אם מקובצים לפי דובר, רואים את כל מה שהוא מסויים אמר אבל מאבדים את הסדר הcronologic של הדיון.

שאלה 4

היהți נשאר עם גישת המשפט הבודד, אבל הייתִ עושה לה שדרוג קטן.
הweeney: לשמור כל משפט בנפרד (כי זה נותן את הגמישות הכי גדולה למשימות שונות), אבל להוסיף שדות שישמרו את ההקשר:

- **קישור למשפט הקודם** - ככה אפשר תמיד לגשת אחרת ולראות את הרצף.
 - **תגיות תוכן** - למשל, תקציר קצר של מה הפרוטוקול עוקב בו או נושא הדיון הספציפי.
- בעצם זו גישה היברידית: מקבלים את יכולת לעבוד ברמת פירוט של משפט בודד, אבל לא מיותרים על האופציה להבין את התמונה הרחבה כשצריך.