עיבוד שפות טבעיות - תרגיל 1

סמסטר א' תשפ"ה

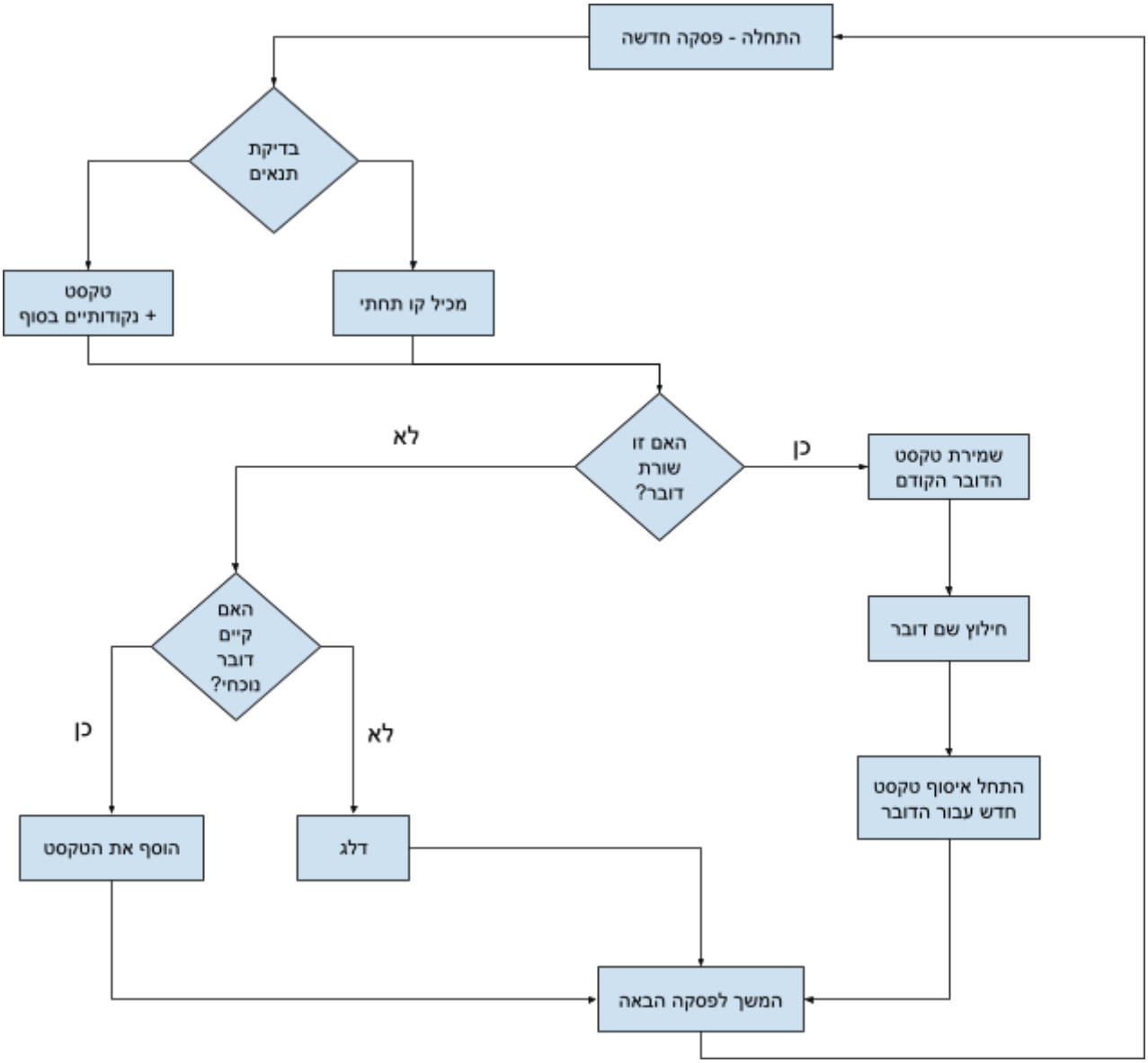
מגישים

שם	ת.ז
אור ישלח	314861584
עודד דואק	039945522

חלק 1 - קוד

3. שליפת טקסט בעל תוכן

ראשית לפי קריאת הפרוטוקולים ראינו כי החלק של הדיון בין הדוברים תמיד מתחיל בדברי היו"ר,
לכן מפסקה זו התחלנו לנתח את הטקסט.
שמנו לב ששמות הדוברים (לפעמים עם תפקיד ומפלגה) מופיעים עם קו תחתון ונקודותיים בסוף ולכן אלה הבדיקות שעשינו כדי לזהות אותם.
כעת לקריאת הפסקאות השתמשנו באלגוריתם הבא:



a) ניקוי שמות דוברים

אם קיים שם המפלגה הוא מופיע בסוגריים ולכן הסרנו את הסוגריים עם התוכן שלהן.
הגדרנו מספר קבוצות של מילים מיוחדות:

- בעלי תואר/תפקיד כמו: {..." 'ד"ר', "פרופ"', 'עו"ד', 'רב', 'ניצב', 'היו\"ר'}
- מילות עצירה כמו: {...'במשרד', 'בממשלה', 'ביטחון', 'פנים', 'לביטחון', 'למשטרה', 'לחקלאות"}
- רשימה של שמות תקינים שמתחילים באות ה' כמו: {...'הגר', 'הדס', 'הדסה', 'הדר'}

בתחילה, האלגוריתם מגדיר מחרוזת ריקה עבור השם (`name = ' '`)
תוספות כמו תפקיד (אם קיימות) מופיעות לפני השם ולכן הפכנו את סדר המילים וכאשר הגענו בלולאה למילה המייצגת תואר או תפקיד אנחנו מפסיקים את הרכבת השם על ידי הפסקת הלולאה.
כדי לדעת אם הגענו לתואר או תפקיד אנחנו בודקים את הדברים הבאים:

- המילה הנוכחית מתחילה באותיות 'וה' (כי הדבר יכול לרמז על מילה אחרונה בתפקיד, למשל שרת התרבות והחינוך)
- המילה הנוכחית מתחילה באות ה אך היא לא שם (כי הדבר יכול לרמז על התפקיד, למשל: שר המשפטים, אך יש לבדוק שמילה זו היא לא שם כמו למשל הרצל)
- המילה הנוכחית היא תואר/תפקיד מתוך רשימה שהגדרנו (למשל 'ד"ר')
- המילה הנוכחית היא ברשימת מילות עצירה שהגדרנו כדי לעצור למשל במילה פנים במקרה הבא: שר לביטחון פנים

בכל מקרה נעצור אם הגענו כבר ל-5 מילים כי לא סביר שיופיעו שמות ארוכים מכך.

- **נראה דוגמת הרצה:**

קלט: סגנית היו"ר במשרד הבריאות ד"ר הדס שטיינר
הרצה:
1. קוראים מהסוף: ["שטיינר", "הדס", "ד"ר","הבריאות", "במשרד", "היו"ר", "סגנית"]
2. סריקה מפורטת:
"שטיינר" -> נכלל ✓ (שם משפחה)
"הדס" -> נכלל ✓ (נמצא ברשימת השמות התקניים שמתחילים ב-"ה")
"ד"ר" -> עצירה (תואר)
פלט: "הדס שטיינר"

b)

בעיות בשימוש בשמות כפי שהופיעו בפרוטוקולים לפני ניקיון השמות:

- חוסר אחידות בייצוג שמות

דוגמאות לאותו אדם:

ח"כ בנימין נתניהו (הליכוד), ראש הממשלה בנימין נתניהו, מר בנימין נתניהו

ללא ניקיון כל אחד עלול להחשב כדובר נפרד. בנוסף, הדבר יכול להקשות על אגרגציה לפי דובר.

- מידע נוסף בסוגריים

לדוגמא: דוד לוי (מחליף את יו"ר הוועדה)

יכול לשמור מידע לא מהותי וגורם לקטלוג לא נכון של התפקידים.

בעיות בשימוש בשמות כפי שהופיעו בפרוטוקולים אחרי נקיון השמות:

- שני דוברים שונים נראים זהים:

לדוגמא:

לפני	אחרי
ח"כ ישראל כהן (ש"ס)	ישראל כהן
ישראל כהן (נציג התאחדות התעשיינים)	ישראל כהן

- איבוד מידע

לדוגמא:

לפני הניקוי: "היועצת המשפטית לוועדה, עו"ד שגית אפיק" אחרי הניקוי: "שגית אפיק"

מידע שאבד: - תפקיד ספציפי בוועדה
- הכשרה מקצועית
- הקשר להליך

איבוד המידע יכול להשפיע על הקונטקסט של הדיון ומשמעות המשפטים.

- ניקוי אגרסיבי מדי

במקרים מסוימים יכול להיות שחלק מהשם של הדובר יימחק.

טקסטים המופיעים באמצע הפרוטוקול

בחרנו לצרף טקסטים אלו (כמו פרטי הצבעה או נושא של הצעת חוק) לדובר האחרון. הסיבות לכך הן:

- שמירה על סדר הכתוב, דבר שיכול לשמור על הקונטקסט.
- פשטות - ייתכן שיהיה קשה לדעת איפה מתחיל ונגמר טקסט שאינו שייך לשום דובר וטעויות עלולות לגרום לרעש.
- מחיקת טקסט כזה יכול לגרום לאיבוד מידע רלוונטי.
- לא כל טקסט כזה שייך ליושב ראש ולכן לא נכון תמיד לצרף אותו אליו.

התעלמנו מקריאות כי צירוף שלהן לדובר האחרון מכניס רעש ויכול אף לסתור את מה שהדובר אמר.

חלוקה למשפטים

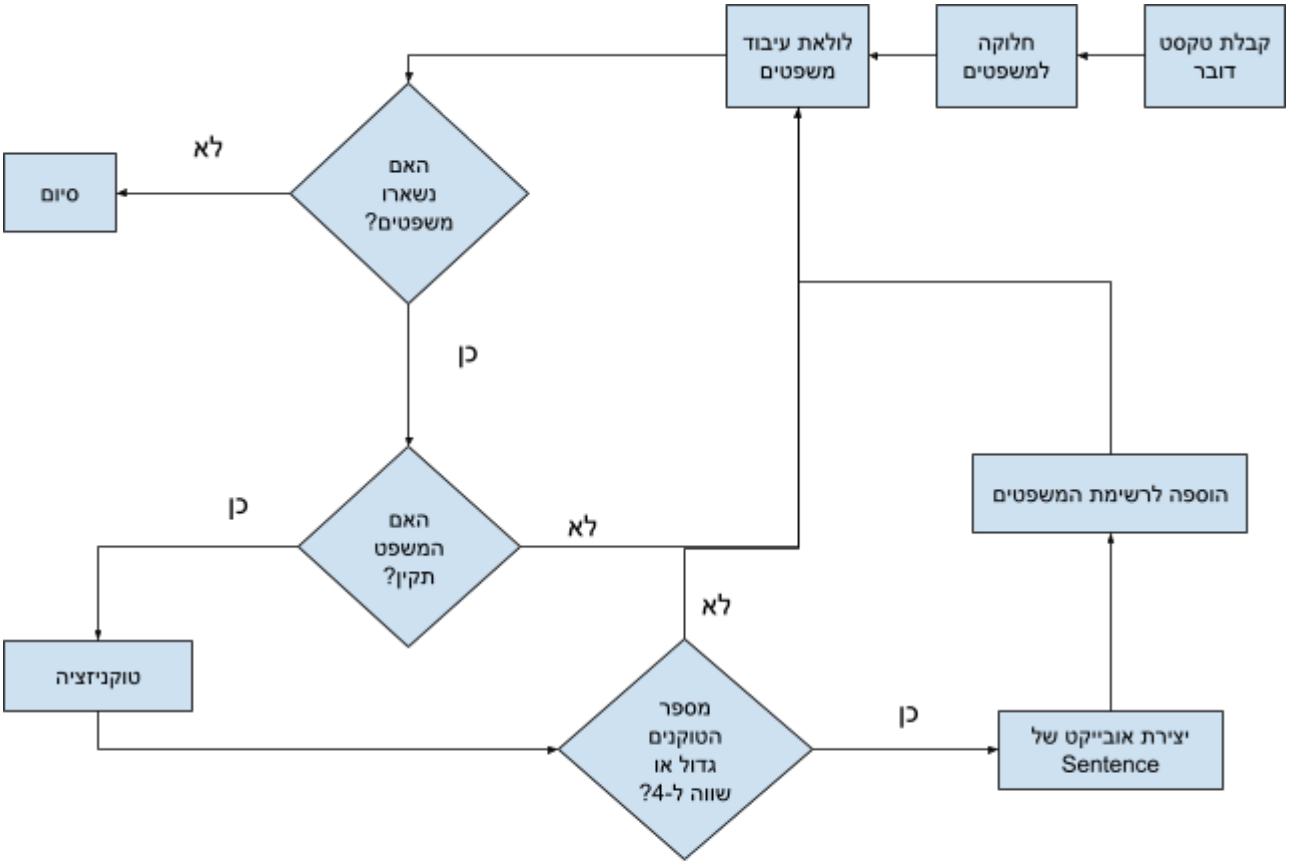
ביצענו את החלוקה למשפטים באמצעות ביטוי רגולרי המזהה סימני פיסוק סופיים (נקודה, סימן שאלה, סימן קריאה)

ואחריהם לפחות רווח אחד ואז טוקן. זה פתר לנו את הבעיות הבאות:

- התמודדת עם ראשי תיבות המכילים נקודות באמצע.
- זיהוי נכון של סוף משפט גם כשיש רווחים מרובים.

ניקיון המשפטים

עבדנו בקריאת המשפטים על פי האלגוריתם הבא:



הגדרנו מספר קריטריונים לסינון משפטים לא תקינים:

- משפטים ללא אותיות בעברית (בדיקה באמצעות regex)
- משפטים המכילים אותיות באנגלית (בדיקה באמצעות regex)
- משפטים קצרים מדי (פחות מ-4 טוקנים)
- משפטים חלקיים המסומנים על ידי תווים דוגמת "---" (בדיקה באמצעות regex לאחר נורמליזציה של המקפים-יש כל מיני פורמטים שונים)

טוקניזציה

פיתחנו ביטוי רגולרי מורכב ובעזרת המשך ניתוח מתבצע הטיפול הבא:

- הפרדת סימני פיסוק לטוקנים נפרדים (לדוגמא: ,;!?)
- התייחסות למילים ומספרים כאל טוקנים
- הפרדת מילים שמכילות גם אותיות וגם מספר לטוקנים נפרדים (כגון: 194א)
- שמירה על ראשי תיבות כיחידה אחת. במקרה שמרכאות מופיעות באמצע מילה אנו מתייחסים לכל המילה והמרכאות כטוקן אחד, אחרת אנו מתייחסים כאל טוקנים נפרדים. במקרה שגרש מופיע באמצע או בסוף מילה אנחנו מתייחסים אליו ואל המילה כטוקן אחד, אחרת כטוקנים נפרדים.
- הפרדת סימנים מיוחדים לטוקנים נפרדים (כגון: %&@)

חלק 2 - שאלות

שאלה 1

ראשית נבחן יתרונות וחסרונות של הפיצול באמצעות דוגמאות במקרים שונים:

🟢 הפיצול מועיל
🔴 הפיצול מזיק
⚠️ מקרה מורכב

קריטריון	דוגמא	הסבר
ניתוח מורפולוגי	לפני: וכשהלכתי אחרי: ו+כש+הלכתי	🟢זיהוי קל של השורש "הלך" והתחיליות
הפחתת אוצר מילים	לפני: ספר, הספר, וספר אחרי: תחילית + ספר	🟢חיסכון במשאבים ע"י שמירת בסיס אחד מאפשר גם חיפוש פשוט במילון
ניתוח רגשות	לפני: וכשראיתי אותו בכיתי אחרי: ו + כש + ראיתי + אותו + בכיתי	🔴פיצול פוגע בהבנת הרצף הרגשי
מילים עם 'ש' בשורש	לפני: שלום אחרי: ש + לום	🔴ה-"ש" היא חלק מהשורש
תחיליות מרובות	לפני: מהבית אחרי: מ+ ה + בית	⚠️יעיל למורפולוגיה אך דורש תשומת לב לקשר בין תחיליות
זיהוי זמנים	לפני: כשאלך אחרי: כש + אלך	🟢מאפשר זיהוי קל של זמן הפועל (עתיד)
הקשר תחבירי	לפני: וכאשר, אבל אחרי: ו + כאשר, לא ניתן לפיצול	🔴פיצול עלול לפגוע בהבנת הקשר בין משפטים ⚠️לא ניתן לפצל מילת קישור בסיסית

לדעתנו, פיצול מוספיות בעברית **צריך להיות מבוסס על מטרת המשימה**.
כשהמטרה היא ניתוח מורפולוגי או הפחתת גודל אוצר המילים - הפיצול יעיל.
לעומת זאת, כשחשוב לשמור על הקשר סמנטי מלא (למשל בניתוח רגשות או הבנת טקסט) - עדיף להימנע מפיצול.

שאלה 2

היינו מציעים את הפיצול הבא: "ו-כש-יבואו", מהסיבות הבאות:

- "ו" - מילת חיבור
- "כש" - מילת קישור (עונה על השאלה 'מתי')
- "יבואו" - הפועל הבסיסי

שאלה 3

🟢 יתרון 🔴 חיסרון

קריטריון	שמירת רשומה לכל פרוטוקול	שמירת כל משפט כרשומה נפרדת
יעילות אחסון	🟢 פחות רשומות במערכת (אחת לכל פרוטוקול)	🔴 יש יותר רשומות עם מידע שחוזר על עצמו, מה שמכביד על האחסון
גמישות בחיפוש	🔴 דורש חיפוש בתוך רשומה ארוכה ומורכבת	🟢 אפשר לסנן ולחפש לפי כל שדה בקלות
יכולת עיבוד	🔴 צריך לפרק את הטקסט לפני העיבוד	🟢אפשר לעבד כל משפט בנפרד או במקביל
שמירה על הקשר	🟢 שומר על ההקשר המלא של הדיון באופן טבעי	🔴 יש צורך בתיעוד נוסף כדי לשמר את הקשר בין משפטים בפרוטוקול
דגימה אקראית	🔴 דורש פירוק הטקסט לפני דגימה	🟢קל לדגום משפטים אקראיים

קריטריון	שמירת רשומה לכל דובר	שמירת כל משפט כרשומה נפרדת
יעילות אחסון	🟢 פחות רשומות במערכת (אחת לכל דובר)	🔴 יש יותר רשומות עם מידע שחוזר על עצמו, מה שמכביד על האחסון
גמישות בחיפוש	🔴 מקשה על חיפוש לפי הקשר או זמן	🟢 אפשר לסנן ולחפש לפי כל שדה בקלות
יכולת עיבוד	🟢מתאים לניתוח דפוסי שפה של דובר ספציפי	🟢אפשר לעבד כל משפט בנפרד או במקביל וניתן לבצע אגרגציה כדי לנתח דפוסי שפה של דובר ספיציפי
שמירה על הקשר	🔴 מאבד את ההקשר הכרונולוגי והרצף של הדיון	🔴 יש צורך בתיעוד נוסף כדי לשמר את הקשר בין משפטים בפרוטוקול
דגימה אקראית	🔴 דורש פירוק הטקסט לפני דגימה	🟢קל לדגום משפטים אקראיים

שאלה 4

נבחר לשמור את המידע בקורפוס באמצעות שמירת כל משפט כרשומה נפרדת כפי שעשינו בתרגיל.
בחרנו זאת מהסיבות הבאות:
גישה זו מאפשרת גמישות גבוהה ל-NLP מאחר וניתן לנתח משפטים בודדים וגם לבצע איחוד לפי פרוטוקולים או דוברים ורק אז לנתח. בנוסף, קל לבצע סינונים לפי פרמטרים שונים.
יתרון נוסף הוא מבחינת יעילות חישובית, כל משפט הוא יחידה עצמאית קטנה ולכן קל יחסית לבצע עיבוד מקבילי.
שיטה זו מאפשרת לבצע חיפושים מדויקים ברמת המשפט ולשלוף מידע ספציפי בלי לטעון פרוטוקולים שלמים.

אומנם קיימים חסרונות לשיטה זו, כמו נפח אחסון גדול יותר וקושי בשמירה על הקשר בין משפטים סמוכים.
אך למרות החסרונות, היתרונות חשובים בהקשר של מחקר טקסטואלי ועיבוד שפה טבעית.