עיבוד שפות טבעיות

תרגיל בית 1

שם: רנים אבראהים (212920896), אסיל נחאס (212245096)

****הבהרה חשובה: הקלטנו כמעט כל שלבי הכתיבה של הקוד וההקלטות נמצאים איתנו אם יש משהו לא מסתדר אז ישנן הסברים בהקלטות שאולי מסבירים מה שנעשה. בכל מקרה הקוד מתועד שורה אחרי שורה.****

שלב 1:

פירוט דרישה 3:

רשום במסמכים הרבה חלקים והיעד שלנו הוא לשלוף טקסטים שנאמרו על ידי דוברים. קודם אנחנו צריכים להבחין בין טקסטים שאנחנו רוצים וטקסטים אחרים, ואז דרך אחת שאנחנו חשבנו עליה היא לשלוף טקסטים שבאים אחרי הסימן " : " כי לרוב המקרים הטקסט שבא אחריה הוא נאמר על ידי דובר, וגם לפעמים מופיע השם של הדובר לפני הסימן ולכן זאת היא תכונה שגם חיפשנו.
דרך המימוש הייתה באופן הבא:

1. **זיהוי דפוסים: השתמשנו בביטויים רגולריים כדי לזהות שורות עם שם הדובר וטקסט שאמר, ושומרים בסוף את שם הדובר ומה שאמר.**
2. **ניקיון שמות דוברים: לפעמים מופיע תפקיד הדובר או שם המפלגה או סימנים אחרים בשם שלו אז מה שעשינו הוא: ניקינו את השם בעזרת ביטויים רגולריים.**
3. **טיפול במקרים חריגים: לפעמים לא מופיע שם הדובר בפסקה ולכן אנחנו מצרפים את זה לדובר האחרון שהופיע ולא בחרנו להתעלם מהם כדי לא לאבד מידע, סיבה אחרת היא שלפעמים גם ממשיכים את הטקסט בלי לזכור את שם הדובר עוד פעם לכן החלטנו לצרף את הדיבור לדובר האחרון כאשר אין דובר חדש שמופיע בטקסט. עוד הסברים למימוש נמצאים בתיעוד הקוד.**

פירוט דרישה 4:

זהינו את הגבולות בין המשפטים על ידי הסימנים ".!?" ולכן כאשר רצינו להפריד בין המשפטים, חיפשנו את הסימנים האלו והוספנו את כל משפט למערך שעשינו. למה בחרנו את השיטה הזאתי? כי פשוט מאוד ברוב המקרים כאשר יש לנו אחד הסימנים זה אומר שהמשפט הסתיים, למרות שזה לא תמיד נכון אבל יש צעדים שונים שעשינו כדי להגיע לרמת דיוק יותר גבוהה. נפרט על הצעדים הנוספים בהמשך.

פירוט דרישה 5:

קודם נתחיל לסנן משפטים לפי שפה בעזרת ביטויים רגולריים, ואחר כך נסנן את המשפטים שאין להם משמעות והם רק מכילים תווים שהם לא אותיות, ואחר כך נסנן את המשפטים שלא שלמים גם בעזרת ביטויים רגולריים.

פירוט דרישה 6:

החלטנו קודם להפריד את המשפטים על ידי רווחים לבנים וגם השתמשנו בביטויים רגולריים כדי לטפל ברוב המקרים החריגים כמו מספרים או סימני פיסוק או מילים שמופרדים על ידי "-" (דאש) וגם הרבה מקרים אחרים שצריך לטפל בהם כטוקן אחד. רוב העבודה בה נעשתה על ידי שימוש בביטויים רגולריים שעזרו לזהות תבניות שהיה צריך לטפל בהן כמו המקרים החריגים שהזכרנו.

שלב 2:

שאלה 1:

פיצול מוספיות בתחיליות של מילים בעברית עשוי להיות מועיל במקרים מסוימים, הדורשות הבנה תחבירית או תחבירית-סמנטית מדויקת, כגון NLP בעיקר במשימות תרגום מכונה, ניתוח תחבירי או סיווג טקסטים. עם זאת, חשוב לקחת בחשבון את החסרונות:

- עומס נתונים.
- מורכבות חישובית.
- סכנת איבוד הקשר סמנטי.

עבור .לכן, ההחלטה האם לפצל מוספיות צריכה להיות תלויה במשימה הספציפית משימות שמטרתן הבנה כללית או ניתוח שטחי של טקסטים, ייתכן ואין צורך לפצל את התחיליות. לעומת זאת, עבור משימות שבהן יש ערך להבנת המשמעות התחבירית, פיצול יכול להועיל.

שאלה 2:

ביצוע טוקניזציה מלאה, שבה המילה "וכשיבואו" מחולקת ל-["ו", "כש", "יבואו"], היא הבחירה המועדפת למשימות עיבוד שפות טבעיות הדורשות הבנה תחבירית וסמנטית מעמיקה. גישה זו מספקת את השקיפות והגמישות הנדרשות תוך שמירה על אפשרות לשחזר את המילה המקורית אם יש צורך

שאלה 3:

יתרונות:

קל יותר לבצע ניתוחים סטטיסטיים או אלגוריתמיים על המשפטים עצמם, כמו .חישוב ממוצע אורך משפטים, זיהוי משפטים חריגים, או סיווג משפטים לפי נושאים

הגישה מאפשרת בקלות לסנן או למיין משפטים לפי פרמטרים כמו דובר, פרוטוקול, .או סוג המשפט

המידע נשמר בפורמט אחיד, המאפשר ניתוח בין-פרוטוקולי ובין-דוברים בצורה פשוטה

כל משפט מכיל את כל המידע הדרוש. זה מונע את הצורך להצליב מידע בין מבנים .שונים בזמן שאילתות

חסרונות:

שמירת משפטים כרשומות נפרדות מגדילה משמעותית את כמות הנתונים, במיוחד .בפרוטוקולים ארוכים עם מאות משפטים

.הדבר עלול לגרום לעלויות אחסון גבוהות יותר ולעומס בזמן עיבוד הנתונים

שמירת משפטים בנפרד מנתקת אותם מההקשר הרחב של הפרוטוקול או הדובר. לעיתים, המשמעות המלאה של המשפט תלויה במיקומו בפרוטוקול או במערכת .היחסים עם משפטים אחרים

אם נרצה לשחזר את המידע ברמת הפרוטוקול המלא או הטקסט המלא של דובר .מסוים, נדרש לעבד מחדש את הנתונים על פי הקשרים שנשמרו

שאלה 4:

הבחירה: שמירת המידע כרשומה לכל משפט עם הקשרים משלימים

הבחירה שלנו היא לשלב את היתרונות של שמירת המידע כרשומה לכל משפט יחד עם שמירת הקשרים רחבים יותר על ידי מבנה נתונים משלים. הפורמט שנבחר ישלב :

1. לכל משפט תהיה רשומה הכוללת את כל המידע :רשומות נפרדות לכל משפט .הרלוונטי (כמו שם הפרוטוקול, שם הדובר, מספר הכנסת, סוג הפרוטוקול, וכדומה)
2. לצד הנתונים ברמת המשפט, נוסיף מבנה :שדה הקשרי בפרוטוקול או בדובר .משלים שכולל את כל המשפטים בפרוטוקול או בדובר מסוים

?למה בחרנו בגישה זו

.אפשר להשתמש בנתוני המשפטים בצורה ישירה למשימות ממוקדות

.השחזור של הקשר רחב יותר מתבצע בקלות דרך הנתונים המשלימים

.מתאים לניתוח ברמת המשפט JSONL פורמט

המשלב פרוטוקולים ודוברים מתאים לניתוח רחב יותר, כגון ניתוח JSON פורמט נאומים של דובר מסוים לאורך כמה פרוטוקולים

.מסכנה:

בחירתנו לשלב את הגישה של שמירת משפטים כיחידות עצמאיות יחד עם מבנה משלים לפרוטוקולים ודוברים מספקת את האיזון המושלם בין גמישות, גרנולריות,

ושימור ההקשר. גישה זו מאפשרת לבצע ניתוחים ממוקדים ברמת המשפט ובמקביל לשחזר את ההקשר הרחב יותר למשימות שדורשות זאת.