תרגיל 5
Parsing

מבוא

בהרצאות למדתן על דקדוק חסר הקשר (Context Free Grammar, או בקצרה CFG), ועל דקדוק חסר הקשר הסתברותי (Probabilistic Context Free Grammar). ראיתן שכל דקדוק חסר הקשר יכול להיכתב בצורה הנורמלית של חומסקי (CNF). בתרגיל זה תממשו את אלגוריתם CKY הנלמד בהרצאה על-מנת לבצע parsing למשפטים ע״פ דקדוק נתון הכתוב בצורה הנורמלית של חומסקי, ותתנסו ב-parsing ידני של משפטים שונים.

חלק א׳: מימוש CKY

מצורף קובץ טקסט ובו Probabilistic Context Free Grammar המתואר באופן הבא:

- בקובץ רשימה של חוקים, כל חוק בשורה נפרדת.
- הסמל התחילי נמצא בראש (צד שמאל) של החוק הראשון.
- כל החוקים שראשם זהה יופיעו זה אחר זה.
- ראש חוק מופרד מגוף החוק בסימן <-
- נון-טרמינלים מתחילים באות גדולה, טרמינלים בקטנה.
- לפני כל חוק תופיע ההסתברות לחוק, מופרדת ע״י רווח.

מצורף קובץ נוסף עם משפטים (לשם פשטות, כל האותיות הן lowercase). כל משפט מופיע בשורה נפרדת, כל טוקן מופרד ע״י רווח.

עליכן ליצור קובץ פלט ובו עץ הגזירה (parse) הסביר ביותר של כל משפט (כלומר, זה שהסתברותו הגבוהה ביותר), יחד עם לוג ההסתברות שלו. שימו לב שלא כל משפט נגזר ע״י הדקדוק שקיבלתן, ובמקרה זה עליכן לציין זאת (ולא לייצר את העץ). בבדיקת התרגילים יתקבלו קבצים שונים של דקדוק הסתברותי (נכתב באותו האופן) ומשפטים שונים.

דוגמאות לפלט: עבור משפט שלא נגזר מהדקדוק, יש לכתוב מתחת למשפט הודעה ברורה:

Sentence: some sentence that cannot derive from the given grammar
*** This sentence is not a member of the language generated by the grammar ***

עבור משפט הנגזר מהדקדוק, על הפלט להיראות כך:

```
Sentence: this flight goes to some city
Parsing:
S
    NP
        D  > this
        N  > flight
    VP
        V  > goes
        PP
            P  > to
            NP
                D  > some
                N  > city
Log Probability: -3.14159265359
```

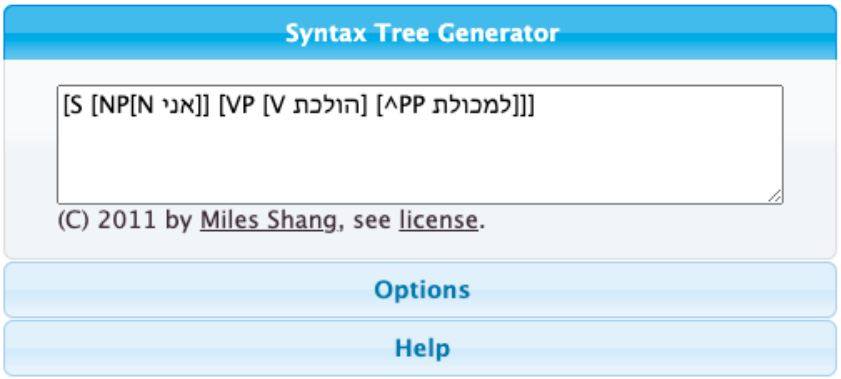
על עץ הגזירה להיכתב באופן מובן (כמו בדוגמא), כלומר, שיהיה ברור מה נגזר מכל חוק. כמו כן, דוגמא זו היא דוגמא כללית, ולאו דווקא תואמת לדקדוק שניתן לכן.

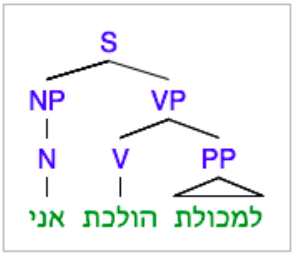
חלק ב': משפטי Garden Path

משפט Garden Path (GP) הוא משפט תקין דקדוקית שעם תחילת קריאתו הקוראת מפרשת אותו באופן מסוים, אך עם המשך קריאתו מתברר כי עליה לפרשו באופן שונה. דוגמא נפוצה בעברית למשפט זה הוא "חולצה מטיילת במדבר": המילה "חולצה" היא דו-משמעית (גם "shirt" אבל גם "was rescued" לנקבה), אך המשמעות הנפוצה יותר שלה היא "shirt". בנוסף, כנראה בעברית פתיחת משפט בשם עצם היא נפוצה יותר מפתיחת משפט בפועל. לכן כשנתחיל לקרוא את המשפט נחשוב שמדובר ב-"shirt". עם זאת, עם קריאת "מטיילת" נתחיל לחשוב שחולצה שמטיילת היא תרחיש פחות סביר. כעת אנחנו מבינות ש"מטיילת" במשפט זה היא לא פועל, אלא שם עצם (נתייחס לבחורה שמטיילת כ"מטיילת"), ו"חולצה" פירושה פה "was rescued".

הדוגמא הנפוצה באנגלית למשפט GP היא "The horse raced past the barn fell". כאשר אנחנו קוראות את המשפט, אנחנו חושבות ש-"raced" הוא פועל עבר, אבל בסופו כשנקרא את המילה "fell", שאינה דו-משמעית (פועל עבר בלבד), נגלה כי "raced" הוא פועל סביל, המשפט הוא משפט משועבד ולכן "raced past the barn" היא פסוקית שעבוד. כלומר, המשפט הוא למעשה "The horse [[that was] race past the barn] fell". ניתן למצוא כאן קריאה נוספת על משפטי GP. כמו כן, בקבוצת הפייסבוק הזו יש לא מעט דוגמאות למשפטים כאלו (ומשפטים משעשעים אחרים).

1. השתמשו באתר הזה על-מנת לחשב את שני עצי הגזירה האפשריים למשפט "חולצה מטיילת במדבר". צרפו לדו"ח את העצים שיצרתן. דוגמא לניתוח משפט בעברית דרך האתר:





לנוחיותכן חוקי דקדוק:

S -> NP VP
NP -> N
NP -> NP PP
VP -> V NP
VP -> V PP

שימו לב שאין צורך לבצע ניתוח רציני, רק לכתוב את העץ באופן שישקף את דו המשמעות. כמו כן, בכפוף לחוקי הדקדוק, המשפט (תחת העץ) יכול להיכתב בסדר שונה: אם הדקדוק מכתיב:

NP -> NP PP

אך המשפט הוא "אסורה הכניסה", אפשר לנתח אותו כ"הכניסה אסורה" כדי שיתאים לדקדוק.

2. חשבו על שני משפטי GP משל עצמכן (באנגלית או בעברית) וכתבו אותם בדו"ח. תארו בדו"ח איפה אפקט ה-GP במשפט, והסבירו מאין הוא נובע. כיצד משפטים כאלו יכולים לאתגר מודלי שפה?
3. חשבו על שני משפטים **דו-משמעיים** (לא בהכרח משפטי GP) באנגלית וכתבו אותם בדו"ח, אחד הנובע מדו-משמעות לקסיקלית (כלומר, דו-משמעות של מילה במשפט), ואחד הנובע מדו-משמעות סינטקטית (כמו "ראיתי את האיש עם המשקפת"). צרו לכל אחד מהם עץ גזירה דרך האתר (אם יש כמה עצי גזירה, צרו את כולם), צרפו תמונה של אלו לדו"ח גם כן. הסבירו (בדו"ח) ממה נובעת הדו-משמעות.

אופן הגשה

1. לתרגיל מצורף קובץ Python בשם **<surname>_<first_name>_ex5.py**. עליכן לממש את התכנית (בגרסה 3.5 ומעלה) ולשנות את שם הקובץ בהתאם (לדוגמא: zeira_yuli_ex5.py). **אין להשתמש בספריות nltk, spacy או transformers**. על הקובץ לרוץ תחת הפקודה:

python <name_ex5.py> <grammar file> <sentences file> <output_filename>

(כמצוין בקובץ עצמו). **לקובץ הפלט יש לכתוב בקידוד utf-8**. הקפידו שזמן הריצה של התכנית לא יעלה על 5 דקות.

2. קובץ PDF בשם **<surname>_<first_name>_report5.pdf** (לדוגמא: zeira_yuli_report5.pdf) ובו דו״ח המפרט על הקוד שכתבתן ועל ההחלטות שקיבלתן במהלך העבודה. על הדו״ח להיות באורך שני עמודים לכל היותר.

*** יש להגיש את שני הקבצים הללו בנפרד, ולא כקובץ zip (ודומיו) ***

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.
יש להגיש את התרגיל עד לתאריך 28.5.22 בשעה 23:59.

בהצלחה!