עיבוד שפות טבעיות סמסטר א' תשפ"א

תרגיל בית מספר 5: **parsing**

מועד הגשת התרגיל: 21.1.2021 בשעה 23:59

מטרת התרגיל היא להבין אלגוריתם probabilistic context free parsing, ואת הקשיים הכרוכים בניתוח תחבירי של שפות טבעיות.
בתרגיל זה אתם מתבקשים לממש את אלגוריתם הניתוח CKY לדקדוקים חסרי הקשר הסתברותיים שנתונים בצורה הנורמלית של חומסקי. עליכם לממש את האלגוריתם בגרסא ההסתברותית שלו על קלט באנגלית. הרחיבו את האלגוריתם שתואר בהרצאה כך שיפיק את עץ הגזירה הסביר ביותר עבור כל משפט בקלט, בנוסף להסתברות של עץ גזירה זה.
אין להשתמש בספריות NLTK או SpaCy או בכל מימוש מוכן אחר של האלגוריתם לתרגיל זה.

היזכרו באלגוריתם הממיר דח"ה כלשהו לצורה הנורמלית של חומסקי (אם אינכם זוכרים, עיינו בספר לימוד לשפות פורמליות או פתחו אלגוריתם כזה בעצמכם). ממשו את האלגוריתם. האלגוריתם כולל שלושה חלקים: ביטול חוקי אפסילון (ניתן להניח שאין כאלה בקלט); ביטול חוקי יחידה (ניתן להניח שאין מעגלים של חוקי יחידה); והמרת חוקים בעלי גוף שכולל יותר משני נון-טרמינלים, או יותר מטרמינל אחד. שלב זה, השלישי, מצריך תוספת סמלים (נון-טרמינלים).

קלט: דח"ה כלשהו (ללא חוקי אפסילון)
פלט: דח"ה שקול בצורה הנורמלית של חומסקי
פורמט הדקדוק (הן קלט והן פלט) :

- רשימה של חוקים, כל חוק בשורה
- הסמל התחילי הוא צד שמאל של החוק הראשון
- כל החוקים שצד שמאל שלהם זהה יופיעו במקובץ
- ראש החוק מופרד מגוף החוק בסימן ->
- נון-טרמינלים מתחילים באות גדולה, טרמינלים בקטנה.

לדוגמה, הדקדוק משקף 558 יינתן כך:

```
NP -> D N
NP -> NP PP
PP -> P NP
D -> the
P -> in
N -> cat
N -> hat
```

ממשו את אלגוריתם הניתוח CYK לדקדוקים חסרי הקשר בצורה הנורמלית של חומסקי. הרחיבו את האלגוריתם שניתן בביתה כך שיפיק גם את עצי הגזירה המושרים על מחרוזות בשפה .
בעת ממשו מנתח תחבירי. הקלט למנתח הוא מחרוזת, והפלט הוא הדפסה של כל עצי הגזירה שהדקדוק משרה על המחרוזת. יתר על כן, עליכם להמיר את עצי הגזירה כך שישקפו את הדקדוק המקורי, לפני ההמרה לצורה הנורמלית. אין צורך לשחזר חוקי יחידה, אבל יש צורך להיפטר בהדפסה מכל הסמלים החדשים שנוספו לדקדוק בעת ההמרה לצורה הנורמלית.

למשל, אם החוק

A -> B C D

הומר לשני החוקים

A -> B NewCD
NewCD -> C D

אזי יש להמיר בפלט גזירות מהצורה:

alpha A beta => alpha B NewCD beta => alpha B C D beta

בגזירות קצרות אך רחבות יותר, מהצורה:

alpha A beta => alpha B C D beta

הדפיסו את עצי הגזירה המתקבלים בפורמט קריא לדוגמה, עץ הגזירה משקף 561 יכול להיות מודפס כך:

(NP (NP (D the)
(N cat))
(PP (P in)
(NP (D the)
(N hat))))

בדו"ח שאתם מגישים פרטו את תיאור האלגוריתמים.

הוראות הגשה:

- הגישו את הקבצים הבאים:
  - קובץ קוד בשם hw5.py.
  - דו"ח בפורמט pdf בשם hw5_report.pdf

- הקוד שמממש את האלגוריתמים ירוץ משורת הפקודה כך:

הפעלה :

python hwf.py <Grammar_in> <Grammar_out> <InputSentences> <OutputTrees>

כאשר:

Grammar_in הוא השם והנתיב של קובץ דקדוק הקלט
Grammar_out הוא השם והנתיב של קובץ דקדוק הפלט,
InputSentences הוא השם והנתיב של קובץ משפטים לניתוח, משפט בשורה
OutputTrees הוא השם והנתיב של קובץ הפלט ובו לכל משפט קלט מודפס המשפט עצמו, ואחריו כל עצי הגזירה שלו, מופרדים בשורת רווח זה מזה.

דוגמאות לקלטים יופיעו בסמוך לתרגיל.

הדו"ח צריך לכלול פירוט של שני חלקי התרגיל כפי שמוסבר בגוף הסעיפים לעיל.
יש להקפיד על עבודה עצמאית, צוות הקורס יתייחס בחומרה להעתקות.
תאריך הגשה 21.1.2021 עד השעה 23:59
שאלות על התרגיל אפשר לשאול בפורום תרגילי בית.
בהצלחה!