Amit Sindani 206902694

Raz Olewsky 315341396

Natural Language Processing – Assignment #3 – Solution

Task #1: using "that" as a subordinate conjunction (connecting sentence clauses)

1. Exploration: make use of the `SelectKBest` function you used in assignment #2 to find words that distinguish between the two sets of identified sentences. What can you say about these markers and peoples' tendency to use (or omit) "that" in certain cases? Explore the data, take some counts from the extracted datasets, and write your reflections in the document you attach to submission.
   Include your code in the version you submit, but don't invoke it in the final version.

   Answer:

נחשב מכל אוסף משפטים שיצרנו (אוסף משפטי explicit ואוסף משפטי implicit) את השכיחות היחסית של כל מילה תיאורית (כמות המופעים של המילה התיאורית באוסף לחלק לכמות המילים באותו האוסף).

להלן טבלה עם 20 המילים התיאוריות שקיבלנו, השכיחות היחסית שלהן והתווית של השכיחות היחסית הגבוהה ביותר לכל מילה:

	Word	Explicit relative freq	Implicit relative freq	Label
0	agree	0.153	0.049	Explicit
1	appears	0.017	0.000	Explicit
2	attack	0.009	0.000	Explicit
3	bet	0.013	0.056	Implicit
4	care	0.074	0.024	Explicit
5	complaining	0.035	0.000	Explicit
6	feel	0.166	0.087	Explicit
7	guess	0.026	0.212	Implicit
8	hope	0.100	0.215	Implicit
9	implies	0.035	0.003	Explicit
10	mean	0.122	0.274	Implicit
11	mention	0.061	0.010	Explicit
12	offended	0.009	0.000	Explicit
13	personal	0.026	0.000	Explicit
14	think	0.450	1.682	Implicit
15	thought	0.096	0.330	Implicit
16	true	0.026	0.007	Explicit
17	understand	0.105	0.035	Explicit
18	willing	0.031	0.000	Explicit
19	wish	0.022	0.156	Implicit

לאחר בחינה של המילים התיאוריות שקיבלנו מהפעלת הפרוצדורה SelectKBest והתבוננות בשכיחות היחסית שלהן, הגענו לתובנה הבאה המתאימה לחלק מהמילים:

עבור מילה תיאורית מסוימת המייצגת פועל – אם השכיחות היחסית שלה באוסף מסוים גבוהה יותר אזי המילה מתארת בצורה יותר מובהקת את האוסף הנ"ל.

בנוסף, השכיחות היחסית הגבוהה יותר משקפת את הנטייה של האנשים להשתמש במילה במשפטים השייכים לאותה קטגוריה האוסף.

<u>לדוגמא:</u>

- השכיחות היחסית של המילים: agree, appears, care, complaining, feel, implies, mention, understand – גבוהה יותר באוסף משפטי explicit לעומת אוסף משפטי implicit.
  נסיק שלאנשים תהיה נטייה גדולה יותר להשתמש במילים אלה כאשר הם עושים שימוש במשפטים מפורשים (משפטים בהם המילה "that" משמשת לחיבור פסוקית ראשית לפסוקית משנה).
- השכיחות היחסית של המילים: bet, guess, hope, mean, think, thought, wish - גבוהה יותר באוסף משפטי implicit לעומת אוסף משפטי explicit.
  נסיק שלאנשים תהיה נטייה גדולה יותר להשתמש במילים אלה כאשר הם עושים שימוש במשפטים מרומזים (משפטים בהם המילה "that" אינה מחברת בין פסוקית ראשית לפסוקית משנה אבל יכולנו להוסיף אותה).

נתייחס כעת גם למילים שאינן נמצאות בקבוצת המילים התיאוריות שקיבלנו.

ישנן מילים שהפרוצדורה SelectKBest תתייחס אליהן כמילים לא תיאוריות.

כשניקח מילה מסוימת מקבוצה זו, נגלה כי השכיחות היחסית שלה באוסף משפטי ה-explicit דומה לשכיחות היחסית שלה באוסף משפטי ה-implicit. כלומר המילה שומרת על יחס מאוזן בין שני סוגי האוספים.

המסקנה היא שישנן מילים שאינן מצביעות על הנטייה של האנשים להשתמש בהן בכל סוג משפט.

מילים אלה "נטרליות" ולא משפיעות על ההחלטה לשימוש במילה "that" בצורה מפורשת בחיבור פסוקית ראשית לפסוקית משנה במשפט או להשמטתה.

<u>הערות לבודקת:</u>

נרצה להסביר את הלוגיקה שלנו לחיפוש המשפטים הרלוונטיים בדאטה סט וסיווגם לקטגוריה הרלוונטית (explicit ו-implicit).

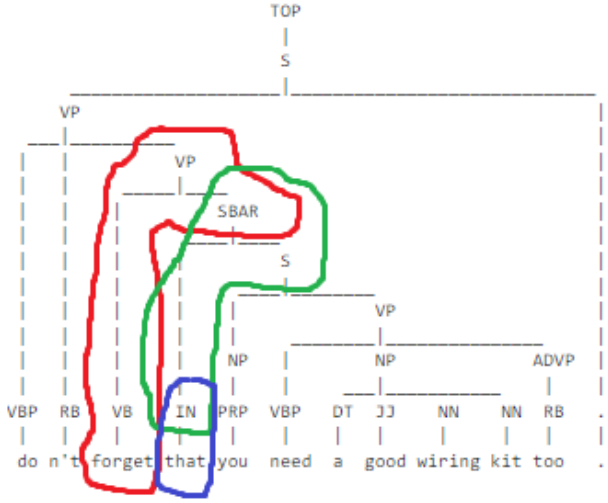
תחילה, חיפשנו בדאטה סט מספר משפטים בהם קיימת המילה "that" בצורה מפורשת ומספר משפטים שניתן היה להוסיף להם את המילה "that".

למשפטים האלה ביצענו ניתוח תחבירי ורצינו להבין האם קיימת חוקיות מסוימת בעצים התחביריים המייצגים משפטים מאותו סוג.

נציג את המסקנות שלנו באמצעות דוגמה על משפט מפורש ומשפט מרומז:

משפט מפורש: "don't forget that you need a good wiring kit too."

נתבונן על הניתוח התחבירי של המשפט המפורש:



<u>המסקנות -</u>

- משפטים מהסוג <u>המפורש</u> מציגים תבנית של חוק הגזירה הבא:

  VP -> (Verb POS such as VB) SBAR

  מסומן באדום

- משפטים אלו מקיימים גם את חוק הגזירה הבא: SBAR -> IN S

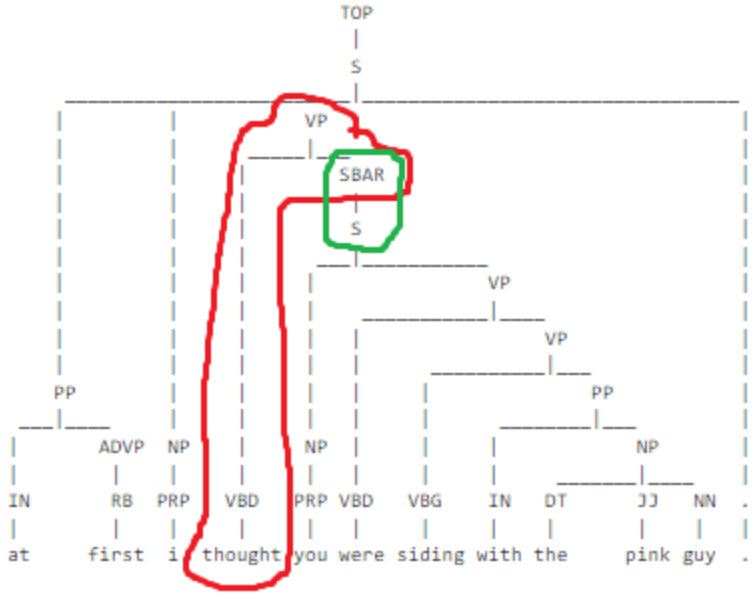
  מסומן בירוק

- המילה that מהווה חלק דיבר של IN (מילת יחס - preposition)

  מסומן בכחול

משפט מרומז : "at first i thought [that] you were siding with the pink guy."

נתבונן על הניתוח התחבירי של המשפט המרומז:



<u>המסקנות -</u>

- משפטים מהסוג <u>המרומז</u> מציגים באופן דומה למשפטים מהסוג המפורש תבנית של חוק הגזירה הבא:

  VP -> (Verb POS such as VB) SBAR

  מסומן באדום

- משפטים אלו מקיימים גם את חוק הגזירה הבא: SBAR -> S

  כלומר פסוקית משנה גוזרת מרכיב דקדוקי יחיד מסוג S (sentence)

  מסומן בירוק

על בסיס המסקנות הנ"ל, כתבנו את הקוד.

הסברים על הקוד שכתבנו מצורפים בעמ' הבא.

בדיקה האם הגענו למרכיב דקדוקי מסוג "VP" ולו בדיוק שני בנים

בדיקה האם חוק הגזירה הבא קיים:

VP → (Verb POS such as VB) SBAR

```
if sub_tree.label() == VP_PHRASE and len(sub_tree) == 2:
    if sub_tree[1].label() == SBAR_PHRASE and sub_tree[0].label() in verb_pos:
        if len(sub_tree[1]) == 2 and sub_tree[1][1].label() == S_PHRASE\
          and sub_tree[1][0].label() == IN_PHRASE and sub_tree[1][0][0] == THAT_PHRASE:
            explicit_set.add(sentence)
        elif len(sub_tree[1]) == 1 and sub_tree[1][0].label() == S_PHRASE:
            implicit_set.add(sentence)
```

בדיקה האם מדובר במשפט שיסווג ל-explicit:

1) נבדוק אם חוק הגזירה הבא קיים:

SBAR -> IN S

2) נבדוק שחלק הדיבר IN מייצג את המילה "that"

בדיקה האם מדובר במשפט שיסווג ל-implicit:

1) נבדוק אם חוק הגזירה הבא קיים: SBAR -> S