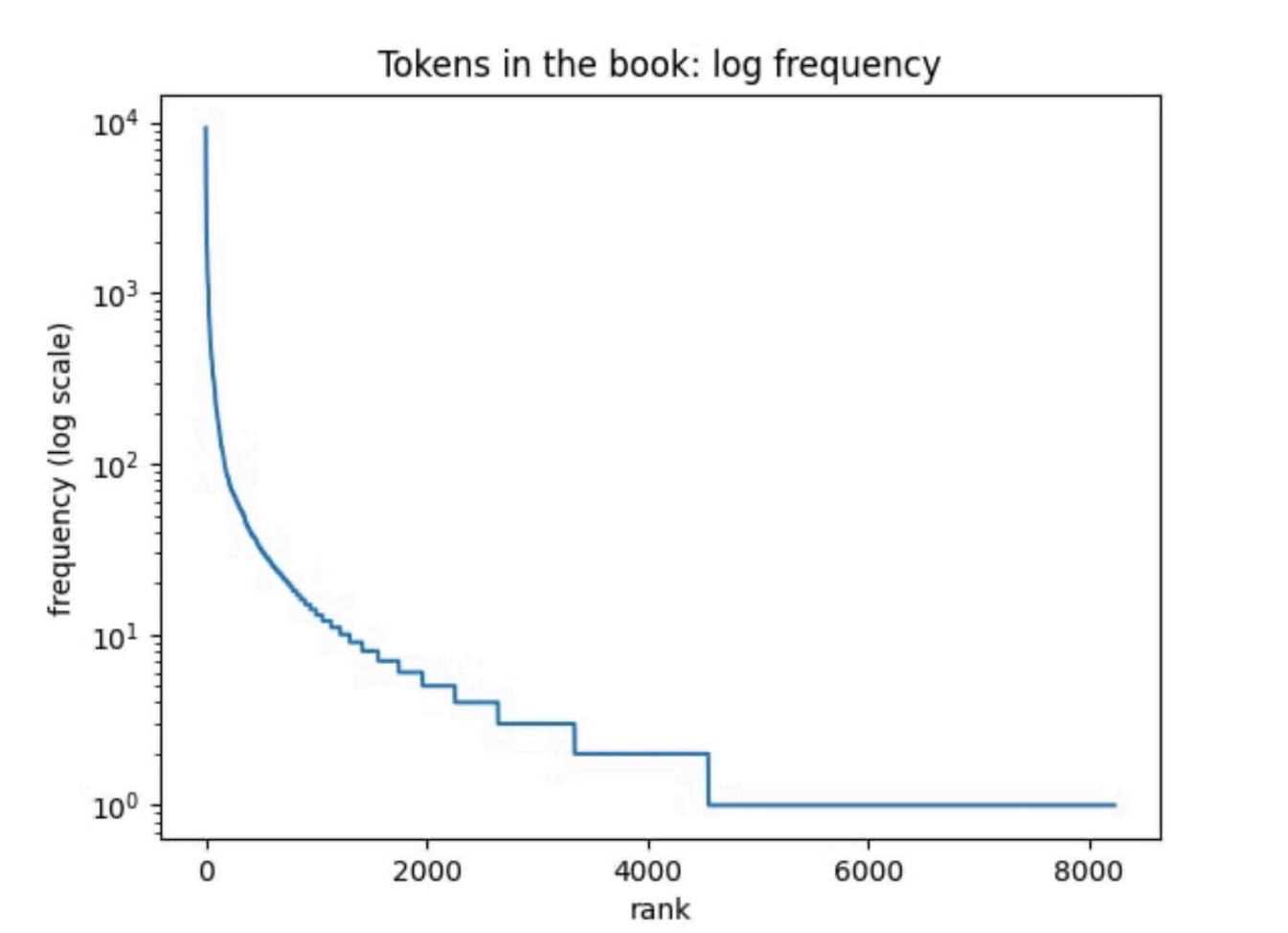
תרגיל 3 - מתוך בעיריות דאטה

מגישים: ענבל משלל - 323839738
שילה דאום - 213571037

(a) שם הספר בו בחרנו הוא "גאווה ודעה קדומה" - Pride and prejudice.

(b) **גרף התוצאה:**



המשך בע"מ הבא ←

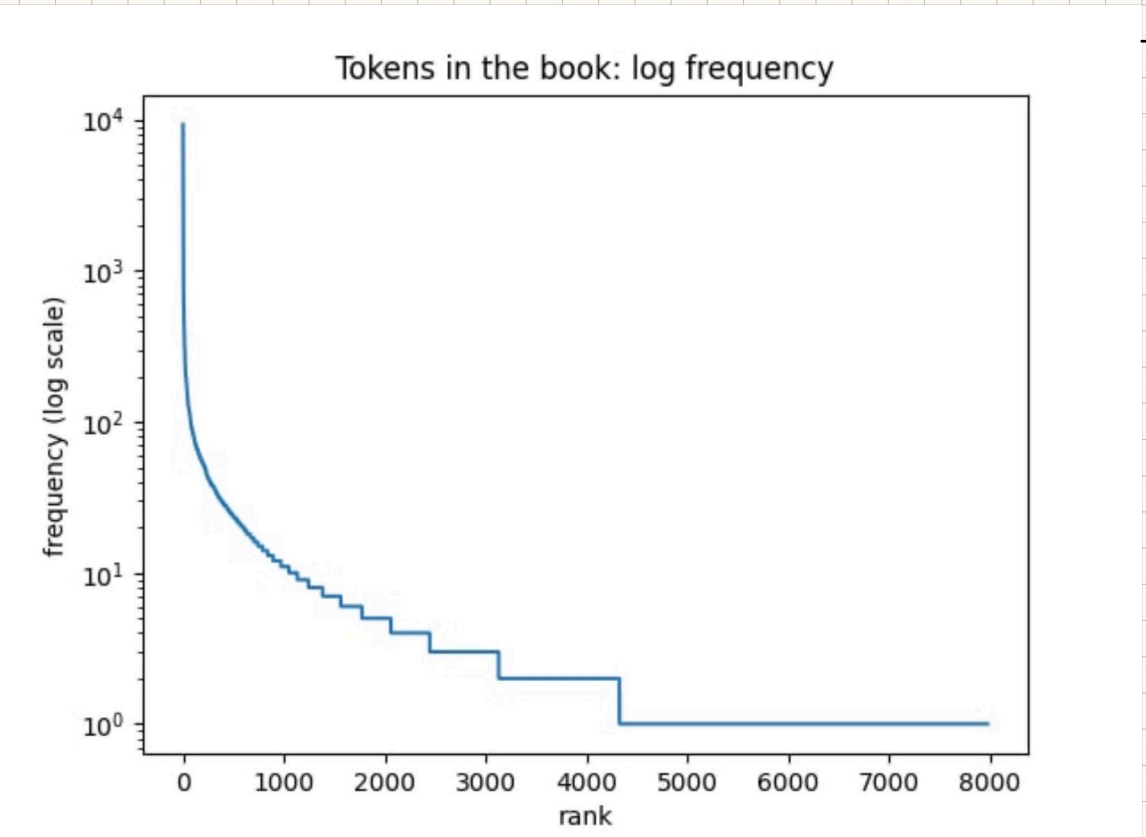
עשרים הטוקנים הנפוצים ביותר וכמות הפעמים שכל אחד

מהם מופיע הם:

- , - 9290
- the - 4221
- . - 4196
- to - 4151
- of - 3694
- and - 3445
- her - 2094
- I - 2029
- a - 1946
- was - 1837
- in - 1832
- “ - 1792
- ” - 1741
- ; - 1553
- not - 1522
- that - 1497
- she - 1363
- it - 1260
- be - 1248
- his - 1174

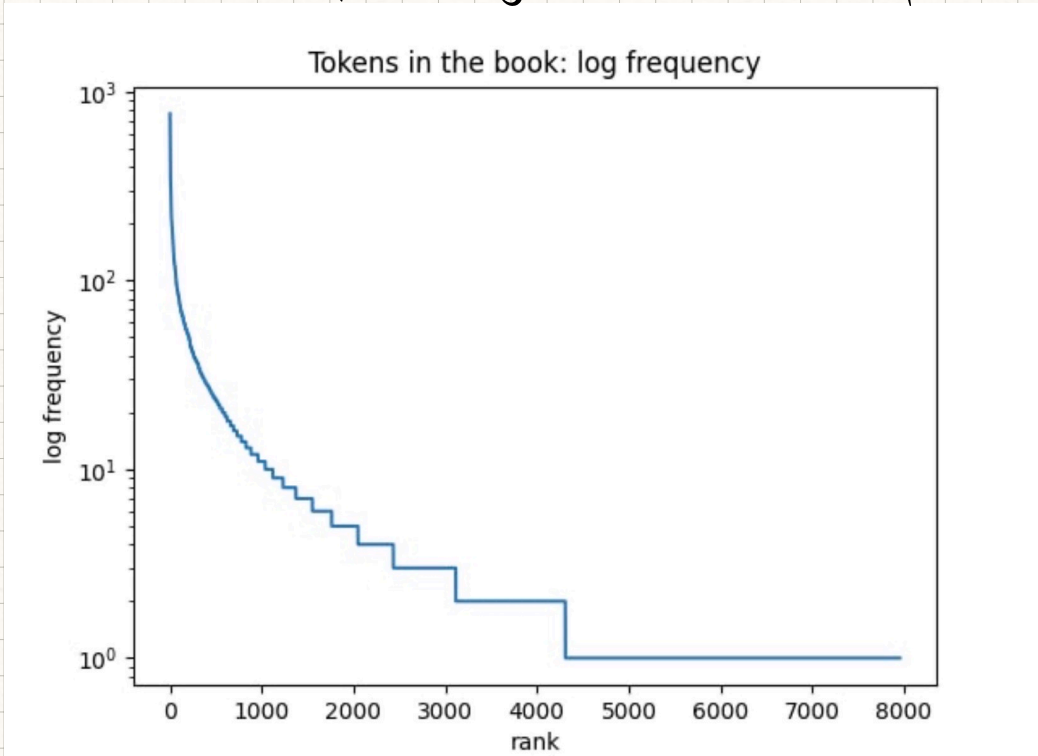
ב) לאחר שמחקנו את ה- Stop-words:

הגרף:



הטוקנים הנפוצים:

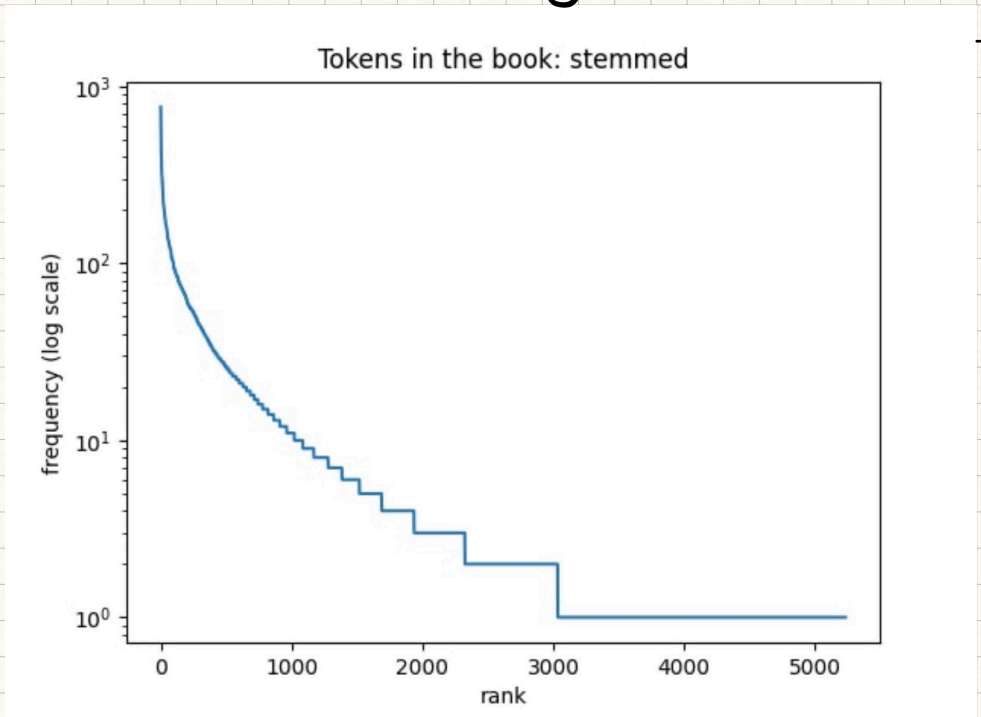
- , - 9290
- . - 4196
- “ - 1792
- ” - 1741
- ; - 1553
- Mr. - 766
- ’ - 723
- Elizabeth - 629
- could - 513
- ! - 495
- ? - 462
- would - 462
- Darcy - 410
- said - 402
- Mrs. - 339
- Bennet - 321
- much - 319
- must - 313
- Bingley - 303
- Jane - 290



- Mr. - 766
- Elizabeth - 629
- could - 513
- would - 462
- Darcy - 410
- said - 402
- Mrs. - 339
- Bennet - 321
- much - 319
- must - 313
- Bingley - 303
- Jane - 290
- Miss - 281
- one - 254
- know - 233
- sister - 214
- never - 212
- soon - 211
- think - 206
- though - 206

d) לאחר שעשינו stemming:

הגרף:



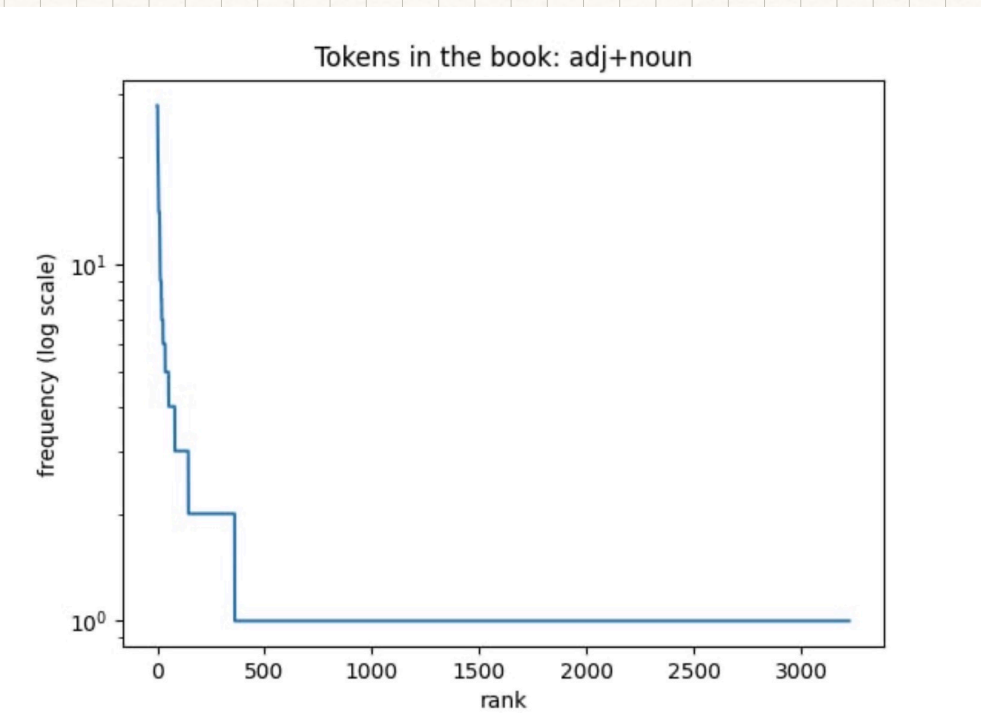
הטוקנים הנפוצים:

- mr. - 766
- elizabeth - 629
- could - 523
- would - 465
- darci - 411
- said - 402
- mrs. - 339
- bennet - 331
- much - 323
- must - 314
- bingley - 309
- jane - 290
- miss - 287
- sister - 287
- one - 274
- know - 269
- ladi - 263
- think - 232
- though - 221
- time - 219

ש) בחלק זה אנו רוצים לנתח את תדירות ההופעות של

הטוויים של שם עצם ותארו:

הגרף:



הטוקנים הנפוצים:

- young man - 28
- Lady Catherine - 28
- great deal - 22
- young ladies - 20
- next morning - 19
- few minutes - 17
- same time - 16
- young lady - 14
- whole party - 14
- next day - 14
- electronic works - 14
- good humour - 13
- electronic work - 11
- last night - 10
- great pleasure - 9
- few moments - 9
- many years - 9
- dear Lizzy - 9
- few days - 9
- younger sisters - 8

(f) דוגמה למשפט שבו נעשה טעות ב- POS tagging:

"... and said in the gentlest of accents..."

המילה gentlest מתארת את המילה accents
(בעבודה הערך הנמוך אבן היא לא סוגה כמילת תואר
אלא כ-מלא. (שם עצם).

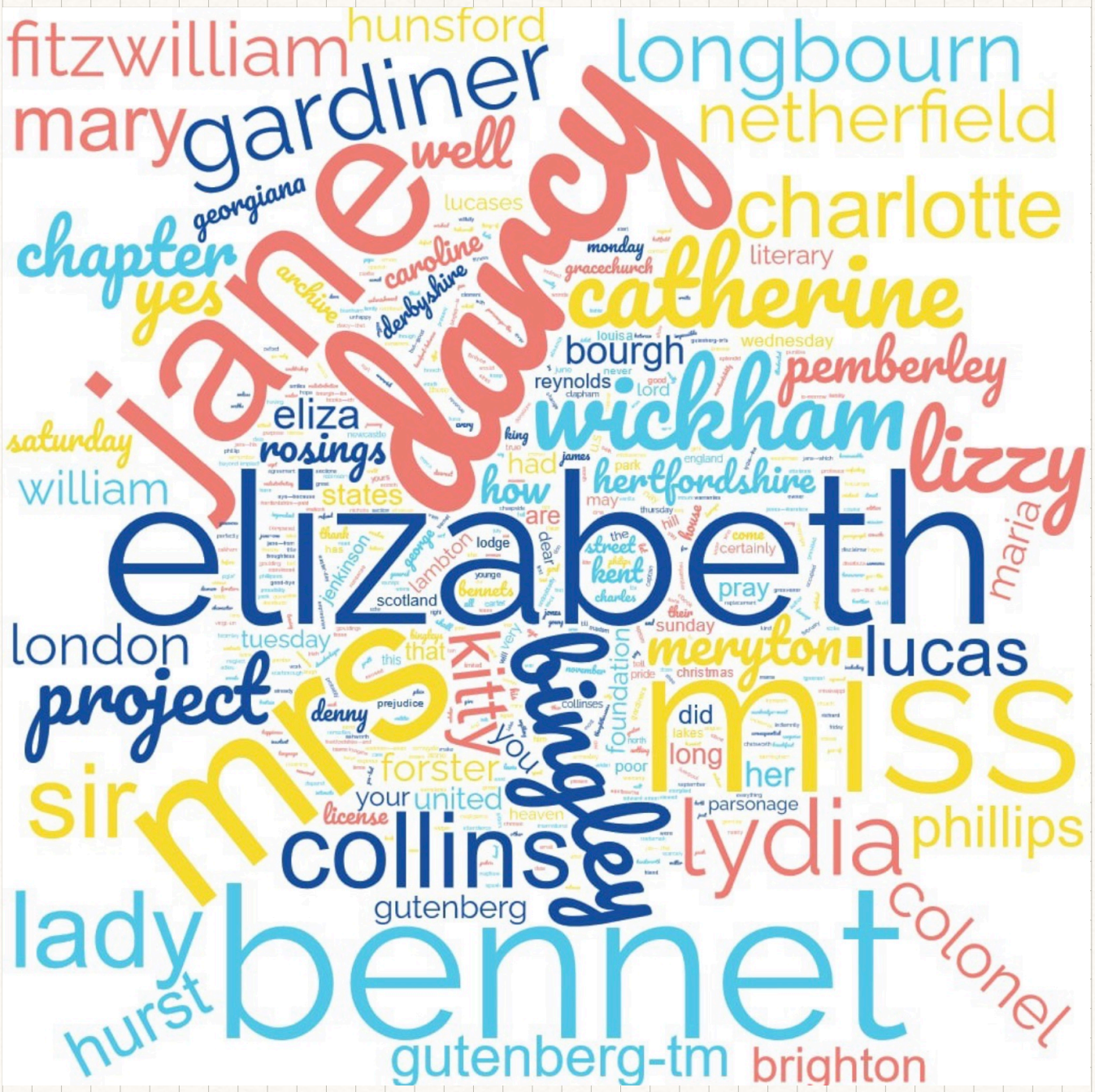
(g) List of POS of the words with the biggest amount of POS:

s - ['NN', 'JJ', 'VBZ', 'VBD', 'POS', 'NNP', 'VBP', 'RB', 'VB', 'FW', 'PRP']
_I_ - ["'''", 'JJ', 'NN', 'VB', 'NNP', 'CC', 'PRP', 'NNS', 'VBP']
that - ['NN', 'JJ', 'NNP', 'VB', 'VBN', 'PRP', 'RB', 'VBG']
till - ['VB', 'IN', 'NN', 'VBP', 'VBZ', 'JJ', 'RB']
ten - ['JJ', 'VBN', 'NN', 'VBZ', 'CD', 'VB', 'NNS']
amongst - ['IN', 'NN', 'VBP', 'JJ', 'VBN', 'VBD', 'RB']
chiefly - ['NN', 'VBN', 'VBD', 'NNS', 'JJ', 'VBP', 'VBZ']
heard - ['VBP', 'VBN', 'RB', 'VBD', 'NN', 'VB']
half - ['VBN', 'NN', 'PDT', 'JJ', 'DT', 'VB']
her - ['RP', 'VB', 'JJ', 'NNP', 'NN', 'NNS']

List of POS of the words with the smallest amount of POS:

- newsletter - ['NN']
- email - ['NN']
- subscribe - ['VB']
- includes - ['VBZ']
- facility - ['NN']
- PG - ['NNP']
- main - ['JJ']
- edition - ['NN']
- volunteer - ['NN']
- network - ['NN']

(h) The Word Cloud:



ניתן לראות כי רוב שמות הדמויות המשמעותיים הם שמות ואכן השמות שמופיעים כגון ג'יין, אליזבת' (שנקראת גם ליזי אליזה), קתרין (שנקראת גם קיטי), מרי ולדיה הן 5 הדמויות הראשיות בסיפור.

i) **ד"ל:** לכתוב ביטוי רגולרי שמוצא את כל המחרוזות עם 2 מלים חוזרות רצופות (לא משנה אם יש בינהן סימני פסוק)

פתרון:

```
regex_expr = r'\b([a-zA-Z]\w*)[\s!\"#$%&\'()*+,.<>=\?@;\-\^_{}~]+\1\b'
```

התוצאות לפי הספר שלנו:

```
had had
I, I
very, very
Lizzy. Lizzy
you, you
it, it
dear, dear
many, many
her her
talents; talents
had had
pride; pride
gratitude;
    gratitude
very, very
where, where
no, no
that, that
thoughtless, thoughtless
you! you
etc., etc
dear, dear
dear, dear
good, good
etc. etc
dear,
    dear
dear, dear
it, it
many, many
```

שאלה 5

המטלה לדרה או שעת.