Ex3 – Submission of practical part

מגישים:

א. אמרי דרור, 305219040, imri.shouach, imri.dror@mail.huji.ac.il

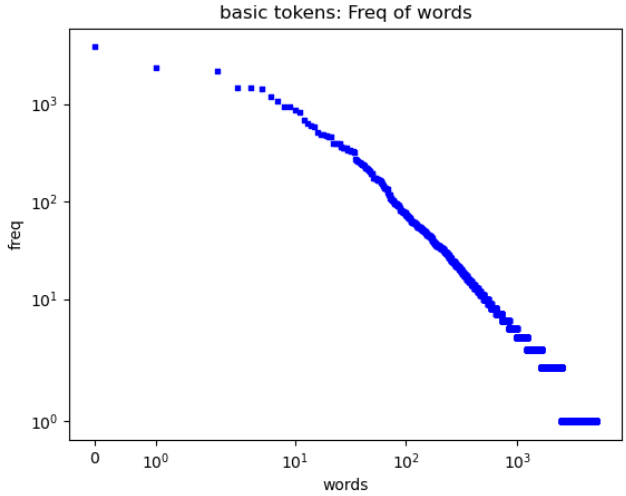
ב. מעיין שרון, 205815566, Maayan.sharon, Maayan.Sharon@mail.huji.ac.il

פתרון:

א. הספר שבו בחנו הוא Peter Pen (בקישור: https://www.gutenberg.org/ebooks/16)

ב. תיוג הטקסט למילים באופן בסיסי (ללא סינון נוסף):

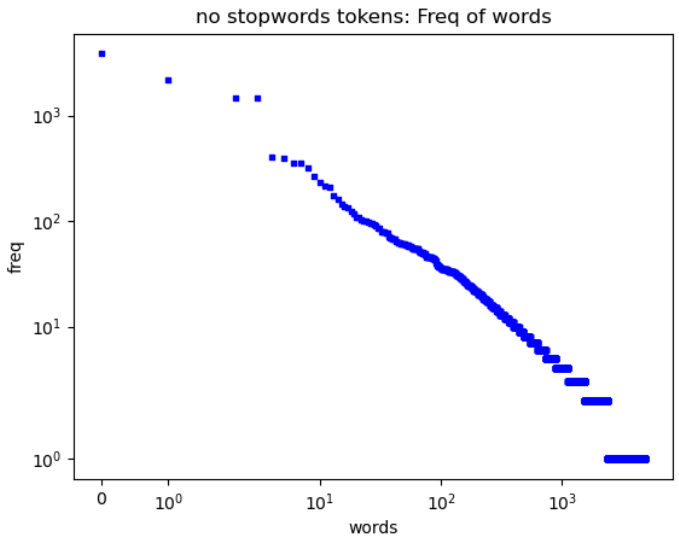
Output:
number of tokens - basic: 5205
top freq words - basic tokens:
[(',', 3837), ('the', 2345), ('.', 2197), ('“', 1467), ('”', 1467), ('and', 1415), ('to', 1182), ('he', 1057), ('was', 934), ('a', 932), ('of', 857), ('it', 816), ('in', 679), ('that', 630), ('she', 602), ('they', 587), ('had', 510), ('you', 481), ('but', 480), ('his', 473)]



ג. תיוג הטקסט למילים לאחר הסרת Stopwords:

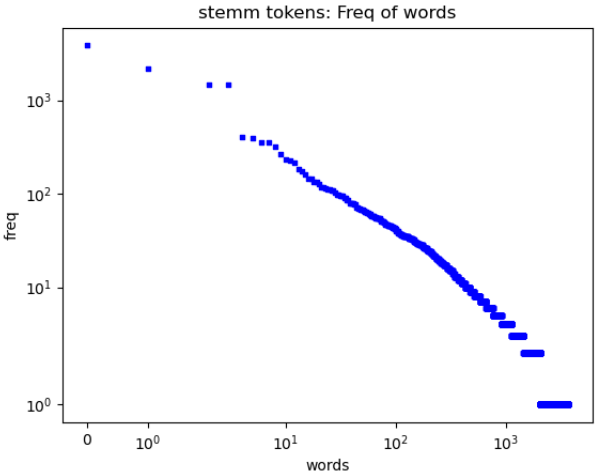
בדקנו עם stopwords נקי, ופעם נוספת אחרי הסרה של סימני פיסוק.

Output:
number of tokens - no stopwords: 4876
top freq words - no stopwords tokens:
[(',', 3837), ('', 2197), ('“', 1467), ('”', 1467), ('peter', 400), (';', 389), ('said', 358), ('wendy', 357), ('?', 319), ("'s", 267), ('!', 235), ('would', 217), ('one', 211), ('hook', 173), ("n't", 159), ('could', 145), ('cried', 136), ('john', 132), ('time', 122), ('darling', 117)]
Without punctuations:
number of tokens - no stopwords: 4864
top freq words - no stopwords tokens:
[('peter', 400), ('said', 358), ('wendy', 357), ("'s", 267), ('would', 217), ('one', 211), ('hook', 173), ("n't", 159), ('could', 145), ('cried', 136), ('john', 132), ('time', 122), ('darling', 117), ('michael', 109), ('see', 107), ('little', 104), ('mother', 101), ('boys', 101), ('children', 98), ('know', 92)]



ד. תיוג הטקסט לפי גזעים (Stems) לאחר הסרת Stopwords :

Output:
number of tokens - stemm: 3710
top freq words - stemm tokens:
[(',', 3837), ('', 2197), ('“', 1467), ('”', 1467), ('peter', 400), (';', 389), ('said', 358), ('wendi', 352), ('?', 319), ("'s", 267), ('!', 235), ('one', 229), ('would', 217), ('cri', 181), ('hook', 174), ("n't", 159), ('could', 145), ('boy', 143), ('time', 133), ('john', 132)]
Without punctuations:
number of tokens - stemm: 3698
top freq words - stemm tokens:
[('peter', 400), ('said', 358), ('wendi', 352), ("'s", 267), ('one', 229), ('would', 217), ('cri', 181), ('hook', 174), ("n't", 159), ('could', 145), ('boy', 143), ('time', 133), ('john', 132), ('look', 126), ('darl', 118), ('mother', 117), ('like', 114), ('go', 112), ('see', 112), ('know', 109)]

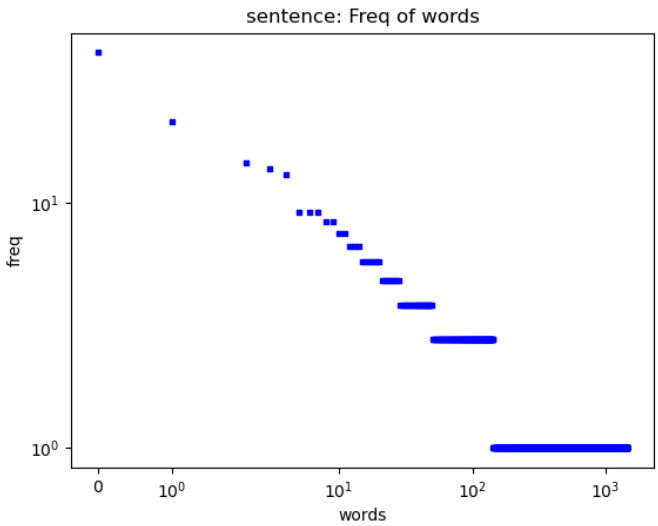


ה. להלן התוצר מתיוג מילים באלגוריתם POS, ובדיקת כמות הופעות של משפטים adj+nn כפי שתואר במשימה :

Output:
number of tokens - sentence: 1677

top freq words - sentence:

[('other boys', 16), ('first time', 8), ('little house', 8), ('good form', 7), ('other hand', 5), ('little boy', 5), ('only man', 5), ('first twin', 5), ('last words', 4), ('long time', 4), ('little girl', 4), ('nursery floor', 3), ('first teeth', 3), ('oh dear', 3), ('little man', 3), ('“ Oh dear', 3), ('old days', 3), ('only one', 3), ('great bed', 3), ('last time', 3)]



ו. להלן משפט מתוך הטקסט, המתויג באמצעות האלגוריתם POS ובו ניתן למצוא טעות בזיהוי :

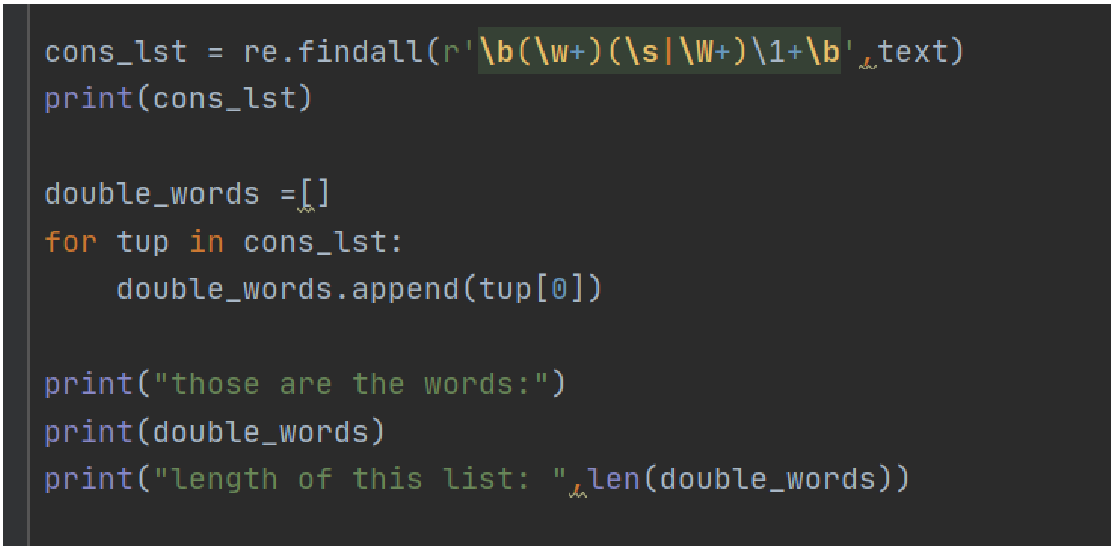
('and', 'CC'), ('the', 'DT'), ('way', 'NN'),('wendy', 'JJ'), ('knew', 'NN'), ('was', 'VBD')

ניתן לראות כי האלגוריתם זיהה את WENDY כסוג של שם תואר פשוט (JJ) על אף שמדובר בשם של ילדה ועל כן אמור להיות סוג של שם עצם (NN).
כמו כן נשים לב שהמילה knew היא פועל בזמן עבר, בעוד תויגה כשם עצם.

ז. ענן מילים על כלל שמות עצם (NNP, NNPS) שזיהה אלגוריתם POS:

ח. זיהוי לפי regular expression:

הרצנו את הקוד הבא:



```
cons_lst = re.findall(r'\b(\w+)(\s|\W+)\1+\b',text)
print(cons_lst)

double_words =[]
for tup in cons_lst:
    double_words.append(tup[0])

print("those are the words:")
print(double_words)
print("length of this list: ",len(double_words))
```

בהתאם לחיפוש שהוגדר, להלן המילים שהתאמתו על הביטוי (כפל מילים זהות) :

First Output (all apparencies) :
those are the words:

['naught', 'nine', 'nine', 'nine', 'pooh', 'had', 'ours', 'had', 'I', 'Nana', 'why', 'George', 'long', 'Wendy', 'tick', 'tick', 'John', 'had', 'tut', 'tut', 'had', 'Latin', 'Pan', 'me', 'Nibs', 'tom', 'tom', 'tom', 'tom', 'tom', 'Hook', 'tap', 'tap', 'Barbecue', 'had', 'tick', 'always', 'Wendy', 'George', 'had', 'Tink', 'Woman']

length of this list: 42

לאחר מכן בדקנו מילים יחידות (כלומר המרנו את הרשימה לset) :

Second Output (unique worsa) :
{'why', 'pooh', 'me', 'Nana', 'nine', 'John', 'Latin', 'ours', 'Tink', 'tom', 'George', 'tut', 'Barbecue', 'tap', 'Hook', 'had', 'naught', 'long', 'always', 'Pan', 'Nibs', 'Woman', 'I', 'tick', 'Wendy'}
Size of this set : 25

כמו כן יש לציין כי לפי ניסוח השאלה, היה צריך ל"תפוס" שתי מילים עוקבות המופרדות ברווח או בסימן פיסוק כלשהו, או כלשון השאלה, עם רווח או סימן פיסוק כלשהו בין שתי המילים העוקבות.

כתוצאה מכך ה regex שבנינו 'תופס' את כל סוגי ה punctuation marks, אך במקרים מסוימים המצב לא מתאר שתי מילים עוקבות באותו משפט, אלא משפט שנגמר במילה מסוימת ולאחר מכן מתחיל משפט חדש עם אותה מילה בדיוק.

דוגמא לכך היא המילה NANA, המופיעה בסוף ובתחילת משפט (מפרידה ביניהם נקודה).

על כן, במידה והיינו רוצים לנתח תופעה זאת באופן מדויק יותר, היינו מתעלמים ממקרים מסוג זה על ידי דיוק ביטוי הregex.