Krit Gupta
UIN: 927001565

WEEK 9 ASSIGNMENT

PYTHON CODE

a. Remove stopwords, apply POS and stemming.

```
/Users/hari/Desktop/class/hw1/venv/bin/python /Users/hari/Desktop/class/hw1/textttt.py
-------------------------------------------------------
---------------------- DOCUMENT1 --------------
-------------------------------------------------------

Document contains a total of 86501  terms.
NN              :14769
IN             :9545
DT              :9016
JJ            :6238
VBD             :6051
PRP            :5127
RB             :4825
NNS             :3673
VB             :3333
CC             :3117
VBN             :2161
TO             :1864
PRP$            :1817
VBG             :1513
VBP            :1364
MD             :1229
VBZ             : 963
CD             : 633
WDT             : 515
WRB             : 444
RP             : 423
WP             : 362
POS            : 335
EX             : 309
JJR            : 177
JJS            : 121
PDT             :  96
RBR             :  90
UH             :  33
NNP             :  30
WP$             :  27
"            :  26
RBS             :  23
(            :  14
)            :  14
FW             :  10
NNPS            :   3
$             :   1
```

Krit Gupta
UIN: 927001565

```
Document contains 40501 terms after removing stop words.

said          : 297
would          : 269
one           : 252
man            : 248
could          : 220
men            : 177
tommy           : 166
upon           : 161
eyes          : 156
like          : 153
...          : 144
n't           : 135
professor      : 131
came           : 120
holtz         : 117
two           : 114
time          : 114
world          : 107
room           : 106
back          : 100


Document contains 40501 terms after stemming.

say           : 360
tommy           : 299
one           : 291
would          : 269
man            : 248
could          : 220
go            : 212
come           : 211
eye           : 206
make           : 204
see           : 179
men            : 177
like          : 173
get           : 171
upon           : 168
know           : 165
look          : 155
professor      : 144
...          : 144
hand           : 142
```

Krit Gupta
UIN: 927001565

```
-------------------------------------------------------
--------------------- DOCUMENT2 ----------------------
-------------------------------------------------------

Document contains a total of 108474  terms.
NN            :20593
IN            :13489
DT            :13367
JJ            :6565
RB            :5097
PRP           :4928
VB            :4498
NNS           :4415
VBZ           :4379
CC            :3492
CD            :3121
VBP           :2578
VBN           :2062
TO            :2062
MD            :1717
VBG           :1697
WRB           :1383
VBD           :1325
PRP$          : 950
RP            : 800
WDT           : 743
JJR           : 482
(             : 296
)             : 296
EX            : 289
WP            : 252
RBR           : 234
PDT           : 163
JJS           : 140
POS           :  82
NNP           :  70
RBS           :  41
$             :  26
FW            :   8
WP$           :   3
UH            :   2
"             :   2
NNPS          :   1
#             :   1

Document contains 48364  terms after removing stop words.

water         : 920
air           : 515
would         : 407
one           : 398
light         : 290
wire          : 267
electricity   : 234
glass         : 211
```

```
fig           : 205
way            : 204
heat          : 198
illustration   : 198
made           : 186
make           : 180
tube          : 168
things         : 167
electric       : 167
see           : 162
hot           : 157
two           : 155


Document contains 48364 terms after stemming.

water         : 922
make           : 694
air           : 518
light         : 461
one           : 437
would          : 407
wire          : 334
go            : 292
get           : 291
experiment     : 274
heat          : 258
thing          : 241
electricity    : 240
put           : 234
use           : 232
see           : 232
glass          : 222
fig           : 214
tube          : 212
way            : 211


-------------------------------------------------------
---------------------- DOCUMENT3 -----------------------
-------------------------------------------------------

Document contains a total of 104764  terms.
NN            :21578
IN           :14152
DT            :13333
JJ           :8526
NNS            :5747
CC            :4482
VBZ            :4406
RB            :4088
CD            :3641
VBN            :3529
VB            :2831
VBP            :2016
```

TO	:1941
PRP	:1840
VBG	:1416
MD	:1207
VBD	: 804
WDT	: 804
PRP$	: 645
WRB	: 471
JJR	: 413
(	: 231
)	: 231
RP	: 198
NNP	: 197
RBR	: 180
JJS	: 152
EX	: 125
WP	: 91
RBS	: 83
PDT	: 81
POS	: 71
$	: 49
WP$	: 22
FW	: 8
"	: 4
UH	: 1

Document contains 50186 terms after removing stop words.

water	: 816
air	: 408
fig	: 340
one	: 327
heat	: 289
illustration	: 247
current	: 232
would	: 195
light	: 189
gas	: 177
force	: 167
upon	: 157
pressure	: 157
may	: 152
used	: 137
two	: 132
motion	: 129
temperature	: 125
small	: 123
made	: 122

Document contains 50186 terms after stemming.

water	: 825
air	: 412
heat	: 388

```
fig           : 365
one            : 348
light          : 322
use            : 263
make             : 262
illustration   : 258
current        : 251
force          : 218
gas            : 216
substance      : 197
would          : 195
form           : 182
color          : 181
time           : 175
great          : 164
sound           : 164
produce        : 164


-------------------------------------------------------
---------------------- DOCUMENT4 -----------------------
-------------------------------------------------------

Document contains a total of 83138  terms.
NN             :12920
IN            :9797
DT             :6687
RB             :6028
JJ            :5778
PRP            :5488
VBD             :4552
VB             :4255
CC             :3263
PRP$            :2815
NNS             :2488
VBN             :2402
TO             :2239
MD              :1772
VBP             :1558
VBG             :1516
VBZ             : 895
CD             : 480
WRB              : 437
WP             : 434
WDT              : 430
POS            : 425
RP            : 261
JJR           : 249
JJS           : 216
PDT            : 201
EX            : 165
RBR            : 160
UH            : 121
RBS            :  90
NNP            :  42
```

Krit Gupta
UIN: 927001565

```
WP$          :  29
(            :  23
)            :  23
"            :  19
FW           :   7
```

Document contains 35456 terms after removing stop words.

```
catherine     : 371
could         : 364
would         : 309
tilney        : 204
one           : 191
must          : 190
said          : 180
never         : 159
well          : 159
time          : 149
room          : 143
isabella      : 142
might         : 138
miss          : 133
every         : 130
good          : 127
general       : 122
brother       : 121
though        : 120
thorpe        : 118
```

Document contains 35456 terms after stemming.

```
catherine     : 485
could         : 364
say           : 327
would         : 309
go            : 239
think         : 234
know          : 223
tilney        : 220
one           : 211
miss          : 207
must          : 190
make          : 185
well          : 184
good          : 175
room          : 172
look          : 171
much          : 170
mrs.          : 168
time          : 167
never         : 159
```

```
-------------------------------------------------------
---------------------- DOCUMENT5 -----------------------
-------------------------------------------------------

Document contains a total of 76232  terms.
NN             :15917
IN            :10707
DT             :9479
VBD            :5275
JJ            :5083
RB             :3885
NNS            :3076
CC             :2719
PRP            :2561
VB             :2329
VBN            :1850
TO             :1583
VBG            :1425
PRP$           :1348
CD             : 921
MD             : 722
VBP            : 662
VBZ            : 512
WRB            : 476
WDT            : 450
RP             : 381
POS            : 270
JJR            : 204
WP             : 176
JJS            : 166
EX             : 158
RBR            : 117
PDT            :  96
NNP            :  54
(             :  31
)             :  31
RBS            :  31
"             :  28
FW             :  19
UH             :  10
$             :   6
WP$            :   2
#             :   2
NNPS           :   1
SYM            :   1

Document contains 35190  terms after removing stop words.

spray          : 518
one            : 281
sea            : 226
would          : 222
day            : 216
island         : 206
could          : 195
```

```
came          : 167
sloop         : 163
wind          : 161
time          : 152
voyage        : 145
good          : 144
ship          : 130
cape          : 128
great         : 127
days          : 121
said          : 112
night         : 108
many          : 108


Document contains 35190 terms after stemming.

spray         : 518
day           : 337
sail          : 322
one           : 312
come          : 276
sea           : 275
island        : 275
make          : 237
would         : 222
wind          : 202
say           : 201
could         : 195
ship          : 185
good          : 173
sloop         : 166
time          : 164
voyage        : 157
go            : 154
great         : 145
find          : 144


--------------------------------------------------------
---------------------- DOCUMENT6 -----------------------
--------------------------------------------------------

Document contains a total of 35073  terms.
NN            :6182
IN            :4342
DT            :3843
JJ            :2752
VBD           :2713
RB            :2041
CC            :1657
PRP           :1441
NNS           :1376
VB            :1123
PRP$          : 846
```

```
VBN        : 840
TO         : 691
VBG        : 641
MD         : 388
VBP        : 367
VBZ        : 278
CD         : 249
"          : 188
WDT        : 176
WRB        : 138
RP         : 129
WP         : 110
EX         :  92
JJR        :  91
POS        :  82
RBR        :  68
PDT        :  54
JJS        :  48
RBS        :  12
(          :   9
)          :   9
WP$        :   5
NNP        :   3
FW         :   2
UH         :   1
```

Document contains 15855 terms after removing stop words.

```
time       : 200
one        : 114
little     : 112
upon       : 110
came       : 105
could      :  93
said       :  89
machine    :  85
saw        :  81
seemed     :  71
man        :  70
like       :  69
thing      :  66
traveller  :  60
white      :  59
would      :  59
world      :  52
still      :  51
felt       :  51
must       :  49
```

Document contains 15855 terms after stemming.

```
time       : 213
come       : 155
one        : 121
```

```
upon          : 113
say          : 112
little        : 112
go           : 103
thing         : 100
could         :  93
machine       :  89
saw          :  89
seem          :  82
look          :  77
hand          :  77
like          :  74
see           :  72
think         :  71
man           :  70
make          :  63
find          :  62


-------------------------------------------------------
---------------------- DOCUMENT7 -----------------------
-------------------------------------------------------

Document contains a total of 80223  terms.
NN            :12820
IN            :7869
DT            :7489
VBD           :6024
PRP           :5806
RB            :5447
JJ            :5030
CC            :3914
VB            :3732
NNS           :2594
VBN           :1757
TO            :1727
PRP$          :1536
VBP           :1501
VBG           :1260
MD            :1239
VBZ           : 952
RP            : 608
WRB           : 479
POS           : 464
CD            : 452
WP            : 333
WDT           : 302
EX            : 242
JJR           : 182
PDT           : 135
JJS           : 129
RBR           :  99
UH            :  95
"            :  52
NNP           :  31
```

Krit Gupta
UIN: 927001565

```
RBS          :  25
WP$          :  21
(            :  18
)            :  18
FW           :  11
NNPS         :   1
```

Document contains 36562 terms after removing stop words.

```
n't          : 623
tom          : 574
said         : 356
would        : 287
'll          : 218
could        : 202
time         : 191
huck         : 179
one          : 178
well         : 152
joe          : 147
boys         : 143
upon         : 141
little       : 139
got          : 137
never        : 131
boy          : 128
two          : 120
came         : 118
away         : 116
```

Document contains 36562 terms after stemming.

```
tom          : 814
n't          : 623
say          : 494
go           : 374
get          : 316
would        : 287
boy          : 284
come         : 282
huck         : 256
ll           : 218
time         : 216
know         : 203
one          : 202
could        : 202
well         : 187
take         : 178
see          : 172
make         : 169
joe          : 166
tell         : 156
```

Krit Gupta
UIN: 927001565

```
-------------------------------------------------------
---------------------- DOCUMENT8 -----------------------
-------------------------------------------------------

Document contains a total of 64515  terms.
NN            :11603
IN           :8608
DT            :7828
VBD            :4762
JJ           :4485
RB            :3282
NNS            :3267
CC            :2994
PRP            :2671
VBN            :1709
VB            :1701
VBG            :1553
PRP$           :1255
TO            :1175
VBP            : 710
MD            : 530
VBZ            : 477
CD            : 467
RP            : 345
WDT            : 325
EX            : 199
WP            : 196
WRB            : 179
POS            : 136
JJR           : 131
PDT            : 105
JJS           :  89
RBR            :  62
(            :  19
)            :  19
RBS            :  13
WP$            :   8
UH            :   6
FW            :   5
NNP            :   4
"           :   2
NNPS           :   2

Document contains 30026  terms after removing stop words.

one           : 184
said          : 166
upon          : 165
martians      : 163
people        : 159
came          : 151
man           : 125
time          : 122
black         : 122
saw           : 118
```

```
could       : 117
men         : 110
little      : 110
road        : 104
would       : 103
brother     : 103
us          : 102
night       : 102
way         : 100
went        :  99


Document contains 30026 terms after stemming.

martian     : 247
come        : 246
go          : 228
one         : 205
say         : 190
upon        : 172
people      : 159
see         : 138
seem        : 134
time        : 133
house       : 130
saw         : 129
man         : 125
black       : 122
could       : 117
thing       : 115
make        : 114
little      : 112
road        : 111
men         : 110

Process finished with exit code 0
```

b. Remove stopwords, do not apply POS, apply stemming

PYTHON CODE

```python
import pandas as pd
import string
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.probability import FreqDist
```

Krit Gupta
UIN: 927001565

```
import matplotlib



# #Download nltk supporting packages
# nltk.download('punkt')
# nltk.download('averaged_perceptron_tagger')
# nltk.download('stopwords')
# nltk.download('wordnet')

#open file

for i in range(1,9):

    print("--------------------------------------------------")
    print("-------------------- DOCUMENT" + str(i) +  " ---------------------")
    print("--------------------------------------------------")


    with open ("T" + str(i) + ".txt", "r") as text_file:
        adoc = text_file.read()


    # Convert to all lower case - required
    #a_discussion = ("%s" %df[0:1]).lower()
    a_discussion = ("%s" %adoc).lower()
    a_discussion = a_discussion.replace('-', ' ')
    a_discussion = a_discussion.replace('_', ' ')
    a_discussion = a_discussion.replace(',', ' ')
    a_discussion = a_discussion.replace("'nt", " not")

    # Tokenize
    tokens = word_tokenize(a_discussion)
    tokens = [word.replace(',', '') for word in tokens]
    tokens = [word for word in tokens if ('*' not in word) and word != "''" and \
              word !="``"]
    # Remove punctuation
    for word in tokens:
        word = re.sub(r'[^\w\d\s]+','',word)

    print("\nDocument contains a total of", len(tokens), " terms.")




    # Remove stop words
    stop = stopwords.words('english') + list(string.punctuation)
    # stop_tokens = [word for word in tagged_tokens if word[0] not in stop]
    # Remove single character words and simple punctuation
    stop_tokens = [word for word in tokens if len(word) > 1]
    # Remove numbers and possive "'s"
    stop_tokens = [word for word in stop_tokens \
                   if (not word[0].replace('.','',1).isnumeric()) and \
                   word[0]!="'s" ]
    print("\nDocument contains", len(stop_tokens), \
                    " terms after removing stop words.\n")
    token_dist = FreqDist(stop_tokens)
    for word, frequency in token_dist.most_common(20):
        print('{:<15s}:{:>4d}'.format(word[0], frequency))


    # Lemmatization - Stemming with POS
    # WordNet Lematization Stems using POS
    stemmer = SnowballStemmer("english")
    wn_tags = {'N':wn.NOUN, 'J':wn.ADJ, 'V':wn.VERB, 'R':wn.ADV}
    wnl = WordNetLemmatizer()
    stemmed_tokens = []
    for token in stop_tokens:
        term = token[0]
        pos  = token[1]
        pos  = pos[0]
        try:
            pos   = wn_tags[pos]
            stemmed_tokens.append(wnl.lemmatize(term, pos=pos))
        except:
            stemmed_tokens.append(stemmer.stem(term))
    print("Document contains", len(stemmed_tokens), "terms after stemming.\n")



    # Word distribution
    #fdist = FreqDist(word for word in stemmed_tokens)
    fdist = FreqDist(stemmed_tokens)
    # Use with Wordnet
    for word, freq in fdist.most_common(20):
        print('{:<15s}:{:>4d}'.format(word, freq))
    # Use with Simple Steming, not with WordNet
    #for word, freq in fdist.most_common(20):
        # print('{:<15s}:{:>4d}'.format(word[0], freq))
    fdist_top = nltk.probability.FreqDist()
    for word, freq in fdist.most_common(20):
        fdist_top[word] = freq
    # fdist_top.plot()
```

RESULTS

```
/Users/hari/Desktop/class/hw1/venv/bin/python /Users/hari/Desktop/class/hw1/textttt.py
------------------------------------------------------
--------------------- DOCUMENT1 ----------------------
------------------------------------------------------

Document contains a total of 86501  terms.

Document contains 77409  terms after removing stop words.

t              :5178
o              :2421
a              :2358
t              :1864
```

```
h            :1390
t            :1165
w             :1136
i            :1118
i            :1035
h            : 779
h            : 590
y            : 586
w             : 574
b            : 554
a            : 543
t            : 526
f            : 495
a            : 459
i            : 457
w             : 395
```

Document contains 77409 terms after stemming.

```
t            :13525
a            :6779
s            :6093
w             :5749
h            :5630
o            :4836
i            :4012
b            :3545
m             :3125
f            :3078
c            :3055
d            :2599
p            :2109
l            :2022
r            :1889
g            :1739
n            :1712
e            :1702
y            : 995
u            : 986
```

---------------------- DOCUMENT2 -----------------------

Document contains a total of 108474 terms.

Document contains 95108 terms after removing stop words.

```
t            :8302
o            :3304
a            :2278
```

```
t             :2056
i             :2056
i             :2045
i             :1848
y              :1485
t             :1115
w               : 920
o              : 797
a              : 745
o              : 710
w               : 697
n              : 666
a              : 647
w               : 625
i             : 608
b              : 568
f             : 549
```

Document contains 95108 terms after stemming.

```
t             :16976
i             :8326
a              :7921
o              :7267
w               :6927
s             :6446
c              :4698
b              :4618
f             :4025
p              :3384
h              :3155
m               :3040
e              :2902
l             :2591
d              :2397
y              :2204
n              :1968
g              :1797
r             :1726
u              :1149
```

---------------------- DOCUMENT3 -----------------------

Document contains a total of 104764 terms.

Document contains 92431 terms after removing stop words.

```
t             :8767
o              :4322
a              :3240
i             :2421
```

```
i             :2206
t             :1941
i             : 901
a              : 900
b              : 874
a              : 857
w               : 816
t             : 785
b              : 777
w               : 722
o              : 581
f             : 551
b              : 533
f             : 529
w               : 525
i             : 512


Document contains 92431 terms after stemming.

t             :15498
a              :9041
i             :8281
o              :7602
s             :6298
w               :5643
c              :5098
b              :4758
f             :4576
p              :3913
m               :3386
d              :2687
h              :2438
r             :2420
l             :2291
e              :2274
n              :1666
g              :1374
u              :1145
v              : 936


---------------------------------------------------------
---------------------- DOCUMENT4 -----------------------
---------------------------------------------------------

Document contains a total of 83138  terms.

Document contains 75327  terms after removing stop words.

t             :3174
o              :2358
a              :2304
t             :2239
h              :1562
```

```
i            :1265
w             :1112
i            :1105
s            :1097
n             :1041
y             : 918
t            : 805
b             : 795
f            : 726
h             : 703
a            : 684
w             : 663
b             : 589
h             : 544
i            : 532


Document contains 75327 terms after stemming.

t            :10525
a            :7491
h             :6331
s            :5814
w             :5441
o             :4699
i            :4442
b             :3742
m             :3681
c            :3153
f            :2872
n             :2467
d             :2258
p             :1988
e            :1853
l            :1583
y             :1491
r            :1473
g             :1345
u             : 718


--------------------------------------------------------
---------------------- DOCUMENT5 -----------------------
--------------------------------------------------------

Document contains a total of 76232  terms.

Document contains 68134  terms after removing stop words.

t            :5833
o             :2370
a            :2121
t            :1583
i            :1370
w             :1335
```

```
o            : 976
f            : 726
t            : 722
i            : 703
a            : 643
h            : 550
m             : 533
w             : 527
s            : 518
f            : 509
a            : 488
a            : 417
h            : 400
b            : 382
```

Document contains 68134 terms after stemming.

```
t            :11678
a             :6548
s            :6240
w             :5247
o             :5242
h             :3508
i            :3402
f            :3251
m             :3210
b             :3193
c             :3143
d             :1872
p             :1809
n             :1786
l            :1641
r            :1457
g             :1273
e             :1079
v             : 540
u             : 513
```

```
-------------------------------------------------------
---------------------- DOCUMENT6 -----------------------
-------------------------------------------------------
```

Document contains a total of 35073 terms.

Document contains 30773 terms after removing stop words.

```
t            :2241
a             :1235
o             :1152
t            : 691
w             : 552
i            : 537
m             : 437
```

```
t            : 433
i            : 418
h             : 355
m             : 281
a            : 261
a            : 238
f            : 217
w             : 215
t            : 200
b             : 185
w             : 158
t            : 149
o             : 137

Document contains 30773 terms after stemming.

t            :5487
a             :3001
s            :2423
w             :2098
o             :2056
m             :1894
i            :1679
h             :1588
f            :1388
b             :1254
c            :1126
p             : 916
d             : 909
l            : 895
e            : 650
n             : 650
r            : 639
g             : 550
'           : 436
u             : 430


--------------------------------------------------------
---------------------- DOCUMENT7 -----------------------
--------------------------------------------------------

Document contains a total of 80223  terms.

Document contains 71318  terms after removing stop words.

t            :3794
a             :3124
t            :1727
o             :1466
i            :1309
h             :1251
w             :1180
t            :1022
i            : 955
```

```
y             : 882
h             : 820
t            : 814
'            : 804
w              : 648
n             : 623
t            : 616
b             : 582
f            : 533
h             : 527
h             : 434


Document contains 71318 terms after stemming.

t            :12241
a             :6910
h             :5774
s            :5523
w              :5480
b             :3707
o             :3584
i            :3458
c             :2468
f            :2442
d             :2337
m              :2331
n             :2329
l            :1817
p             :1700
g             :1630
'            :1519
r            :1266
y             :1229
e             :1158


--------------------------------------------------------
---------------------- DOCUMENT8 -----------------------
--------------------------------------------------------

Document contains a total of 64515  terms.

Document contains 57917  terms after removing stop words.

t            :4794
a             :2503
o             :2301
t            :1175
i            :1001
w              : 854
t            : 793
i            : 688
h             : 582
m              : 472
```

```
a            : 451
w             : 449
a            : 444
h            : 427
o            : 378
w             : 369
f            : 348
t            : 342
f            : 327
b            : 294


Document contains 57917 terms after stemming.

t           :10441
a            :5894
s           :4657
o            :4201
w             :4020
h            :3446
i           :2876
m             :2840
b            :2760
f           :2390
c            :2353
d            :1778
p            :1722
r           :1438
l           :1391
n            :1226
e            :1137
g            :1089
u            : 827
v            : 421

Process finished with exit code 0
```

c. Remove stopwords, do not apply POS, do not stem

```python
import pandas as pd
import string
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.probability import FreqDist
import matplotlib
```

Krit Gupta
UIN: 927001565

```
# #Download nltk supporting packages
# nltk.download('punkt')
# nltk.download('averaged_perceptron_tagger')
# nltk.download('stopwords')
# nltk.download('wordnet')

#open file

for i in range(1,9):

    print("------------------------------------------------------------")
    print("----------------------- DOCUMENT" + str(i) +  " -----------------------")
    print("------------------------------------------------------------")


    with open ("T" + str(i) + ".txt", "r") as text_file:
        adoc = text_file.read()


    # Convert to all lower case - required
    #a_discussion = ("%s" %df[0:1]).lower()
    a_discussion = ("%s" %adoc).lower()
    a_discussion = a_discussion.replace('-', ' ')
    a_discussion = a_discussion.replace('_', ' ')
    a_discussion = a_discussion.replace(',', ' ')
    a_discussion = a_discussion.replace("nt", " not")

    # Tokenize
    tokens = word_tokenize(a_discussion)
    tokens = [word.replace(',', '') for word in tokens]
    tokens = [word for word in tokens if ('*' not in word) and word != "''" and \
          word !="``"]
    # Remove punctuation
    for word in tokens:
        word = re.sub(r'[^\w\d\s]+','',word)

    print("\nDocument contains a total of", len(tokens), " terms.")



    # Remove stop words
    stop = stopwords.words('english') + list(string.punctuation)
    # stop_tokens = [word for word in tagged_tokens if word[0] not in stop]
    # Remove single character words and simple punctuation
    stop_tokens = [word for word in tokens if len(word) > 1]
    # Remove numbers and possive "'s"
    stop_tokens = [word for word in stop_tokens \
                   if (not word[0].replace('.','',1).isnumeric()) and \
                   word[0]!="'s" ]
    print("\nDocument contains", len(stop_tokens), \
                        " terms after removing stop words.\n")
    token_dist = FreqDist(stop_tokens)

    for word, frequency in token_dist.most_common(20):
        print('{:<15s}:{:>4d}'.format(word[0], frequency))




    # Word distribution
    #fdist = FreqDist(word for word in stemmed_tokens)
    fdist = FreqDist(stop_tokens)
    # Use with Wordnet
    for word, freq in fdist.most_common(20):
        print('{:<15s}:{:>4d}'.format(word, freq))
    # Use with Simple Steming, not with WordNet
    #for word, freq in fdist.most_common(20):
        # print('{:<15s}:{:>4d}'.format(word[0], freq))
    fdist_top = nltk.probability.FreqDist()
    for word, freq in fdist.most_common(20):
        fdist_top[word] = freq
    # fdist_top.plot()
```

RESULTS:

```
/Users/hari/Desktop/class/hw1/venv/bin/python /Users/hari/Desktop/class/hw1/textttt.py

------------------------------------------------------
--------------------- DOCUMENT1 ----------------------
------------------------------------------------------

Document contains a total of 86501  terms.

Document contains 77409  terms after removing stop words.

t              :5178
o              :2421
a              :2358
t              :1864
h              :1390
t              :1165
w              :1136
i              :1118
i              :1035
h              : 779
h              : 590
y              : 586
w              : 574
b              : 554
a              : 543
t              : 526
```

```
f            : 495
a             : 459
i            : 457
w              : 395
the           :5178
of           :2421
and           :2358
to           :1864
he            :1390
that          :1165
was           :1136
in           :1118
it           :1035
his           : 779
had            : 590
you            : 586
with           : 574
but            : 554
as           : 543
they           : 526
for           : 495
at           : 459
is           : 457
we             : 395


------------------------------------------------------
--------------------- DOCUMENT2 ----------------------
------------------------------------------------------

Document contains a total of 108474  terms.

Document contains 95108  terms after removing stop words.

t            :8302
o             :3304
a             :2278
t            :2056
i            :2056
i            :2045
i            :1848
y             :1485
t            :1115
w              : 920
o             : 797
a             : 745
o             : 710
w              : 697
n             : 666
a             : 647
w              : 625
i            : 608
b             : 568
f            : 549
the           :8302
```

```
of             :3304
and             :2278
to             :2056
is             :2056
it            :2045
in             :1848
you             :1485
that           :1115
water           : 920
on             : 797
as             : 745
or             : 710
when            : 697
not            : 666
are            : 647
with            : 625
if            : 608
by             : 568
from            : 549


--------------------------------------------------------
---------------------- DOCUMENT3 -----------------------
--------------------------------------------------------

Document contains a total of 104764  terms.

Document contains 92431  terms after removing stop words.

t             :8767
o              :4322
a              :3240
i             :2421
i             :2206
t             :1941
i             : 901
a              : 900
b              : 874
a              : 857
w              : 816
t             : 785
b              : 777
w              : 722
o              : 581
f             : 551
b              : 533
f             : 529
w              : 525
i             : 512
the            :8767
of             :4322
and             :3240
is             :2421
in             :2206
to             :1941
```

```
it           : 901
as            : 900
by             : 874
are           : 857
water          : 816
that          : 785
be            : 777
which          : 722
or            : 581
from          : 551
but           : 533
for           : 529
with          : 525
if           : 512


-----------------------------------------------------
--------------------- DOCUMENT4 ----------------------
-----------------------------------------------------

Document contains a total of 83138  terms.

Document contains 75327  terms after removing stop words.

t            :3174
o             :2358
a             :2304
t            :2239
h             :1562
i            :1265
w             :1112
i            :1105
s            :1097
n             :1041
y             : 918
t            : 805
b             : 795
f            : 726
h             : 703
a             : 684
w             : 663
b             : 589
h             : 544
i            : 532
the           :3174
of            :2358
and            :2304
to            :2239
her           :1562
in            :1265
was            :1112
it           :1105
she           :1097
not           :1041
you            : 918
```

```
that        : 805
be          : 795
for         : 726
had          : 703
as          : 684
with         : 663
but         : 589
he          : 544
is          : 532


--------------------------------------------------------
---------------------- DOCUMENT5 -----------------------
--------------------------------------------------------

Document contains a total of 76232  terms.

Document contains 68134  terms after removing stop words.

t           :5833
o            :2370
a            :2121
t           :1583
i           :1370
w             :1335
o            : 976
f           : 726
t           : 722
i           : 703
a            : 643
h            : 550
m             : 533
w             : 527
s           : 518
f           : 509
a            : 488
a            : 417
h            : 400
b            : 382
the           :5833
of           :2370
and           :2121
to           :1583
in           :1370
was           :1335
on           : 976
for          : 726
that         : 722
it          : 703
at          : 643
had          : 550
my           : 533
with         : 527
spray         : 518
from          : 509
```

```
as           : 488
all          : 417
her           : 400
but           : 382


-----------------------------------------------------
--------------------- DOCUMENT6 ----------------------
-----------------------------------------------------

Document contains a total of 35073  terms.

Document contains 30773  terms after removing stop words.

t            :2241
a             :1235
o             :1152
t            : 691
w             : 552
i            : 537
m             : 437
t            : 433
i            : 418
h             : 355
m             : 281
a             : 261
a             : 238
f            : 217
w             : 215
t            : 200
b             : 185
w             : 158
t            : 149
o             : 137
the           :2241
and            :1235
of            :1152
to            : 691
was            : 552
in            : 537
my             : 437
that          : 433
it           : 418
had            : 355
me             : 281
as            : 261
at            : 238
for           : 217
with           : 215
time           : 200
but           : 185
were           : 158
this          : 149
on            : 137
```

Krit Gupta
UIN: 927001565

```
--------------------------------------------------------
---------------------- DOCUMENT7 -----------------------
--------------------------------------------------------

Document contains a total of 80223  terms.

Document contains 71318  terms after removing stop words.

t            :3794
a             :3124
t            :1727
o             :1466
i            :1309
h             :1251
w              :1180
t            :1022
i            : 955
y             : 882
h             : 820
t            : 814
'           : 804
w              : 648
n             : 623
t            : 616
b             : 582
f            : 533
h             : 527
h             : 434
the           :3794
and            :3124
to            :1727
of            :1466
it           :1309
he            :1251
was            :1180
that          :1022
in            : 955
you            : 882
his           : 820
tom            : 814
's           : 804
with           : 648
n't           : 623
they           : 616
but           : 582
for           : 533
had            : 527
him            : 434
```

Krit Gupta
UIN: 927001565

```
-----------------------------------------------------
--------------------- DOCUMENT8 ----------------------
-----------------------------------------------------

Document contains a total of 64515  terms.

Document contains 57917  terms after removing stop words.

t           :4794
a            :2503
o            :2301
t           :1175
i           :1001
w             : 854
t           : 793
i           : 688
h            : 582
m             : 472
a            : 451
w             : 449
a            : 444
h            : 427
o            : 378
w             : 369
f           : 348
t           : 342
f           : 327
b            : 294
the          :4794
and           :2503
of           :2301
to           :1175
in           :1001
was           : 854
that         : 793
it          : 688
had           : 582
my            : 472
as           : 451
with          : 449
at           : 444
he           : 427
on           : 378
were          : 369
for          : 348
they          : 342
from          : 327
but          : 294

Process finished with exit code 0
```

Krit Gupta
UIN: 927001565

d. Do not remove stopwords, do not apply POS, do not stem

```
import pandas as pd
import string
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.probability import FreqDist
import matplotlib



# #Download nltk supporting packages
# nltk.download('punkt')
# nltk.download('averaged_perceptron_tagger')
# nltk.download('stopwords')
# nltk.download('wordnet')

#open file

for i in range(1,9):

    print("------------------------------------------------------------")
    print("----------------------- DOCUMENT" + str(i) +  " -----------------------")
    print("------------------------------------------------------------")


    with open ("T" + str(i) + ".txt", "r") as text_file:
        adoc = text_file.read()


    # Convert to all lower case - required
    #a_discussion = ("%s" %df[0:1]).lower()
    a_discussion = ("%s" %adoc).lower()
    a_discussion = a_discussion.replace('-', ' ')
    a_discussion = a_discussion.replace('_', ' ')
    a_discussion = a_discussion.replace(',', ' ')
    a_discussion = a_discussion.replace("'nt", " not")

    # Tokenize
    tokens = word_tokenize(a_discussion)
    tokens = [word.replace(',', '') for word in tokens]
    tokens = [word for word in tokens if ('*' not in word) and word != "''" and \
              word !="``"]
    # Remove punctuation
    for word in tokens:
        word = re.sub(r'[^\w\d\s]+',"",word)
```

Krit Gupta
UIN: 927001565

```
print("\nDocument contains a total of", len(tokens), " terms.")

token_dist = FreqDist(tokens)

for word, frequency in token_dist.most_common(20):
    print('{:<15s}:{:>4d}'.format(word[0], frequency))





# Word distribution
#fdist = FreqDist(word for word in stemmed_tokens)
fdist = FreqDist(tokens)
# Use with Wordnet
for word, freq in fdist.most_common(20):
    print('{:<15s}:{:>4d}'.format(word, freq))
# Use with Simple Steming, not with WordNet
#for word, freq in fdist.most_common(20):
    # print('{:<15s}:{:>4d}'.format(word[0], freq))
fdist_top = nltk.probability.FreqDist()
for word, freq in fdist.most_common(20):
    fdist_top[word] = freq
# fdist_top.plot()
```

RESULTS:

```
/Users/hari/Desktop/class/hw1/venv/bin/python /Users/hari/Desktop/class/hw1/textttt.py
-------------------------------------------------------
---------------------- DOCUMENT1 -----------------------
-------------------------------------------------------

Document contains a total of 86501  terms.
t              :5178
.              :4818
o              :2421
a              :2358
a              :1968
t              :1864
h              :1390
t              :1165
w              :1136
i              :1118
i              :1035
i              : 821
h              : 779
h              : 590
```

```
y              : 586
w               : 574
b              : 554
a              : 543
t             : 526
f             : 495
the            :5178
.             :4818
of             :2421
and             :2358
a             :1968
to             :1864
he             :1390
that           :1165
was             :1136
in             :1118
it            :1035
i             : 821
his            : 779
had             : 590
you             : 586
with            : 574
but            : 554
as             : 543
they            : 526
for            : 495


----------------------------------------------------------
---------------------- DOCUMENT2 -----------------------
----------------------------------------------------------

Document contains a total of 108474  terms.
t             :8302
.             :4935
o              :3304
a              :2772
a              :2278
t             :2056
i             :2056
i             :2045
i             :1848
y              :1485
t             :1115
;             : 979
w               : 920
o              : 797
a              : 745
o              : 710
w               : 697
n              : 666
a              : 647
w               : 625
the            :8302
.             :4935
```

```
of             :3304
a             :2772
and            :2278
to             :2056
is             :2056
it            :2045
in             :1848
you            :1485
that           :1115
;             : 979
water          : 920
on             : 797
as             : 745
or             : 710
when           : 697
not            : 666
are            : 647
with           : 625


--------------------------------------------------------
---------------------- DOCUMENT3 -----------------------
--------------------------------------------------------

Document contains a total of 104764  terms.
t             :8767
o              :4322
.             :4228
a              :3240
a              :2719
i             :2421
i             :2206
t             :1941
i             : 901
a              : 900
b              : 874
a              : 857
w               : 816
t             : 785
b              : 777
w               : 722
;             : 633
o              : 581
f             : 551
b              : 533
the            :8767
of             :4322
.             :4228
and            :3240
a              :2719
is             :2421
in             :2206
to             :1941
it            : 901
as             : 900
```

```
by            : 874
are           : 857
water          : 816
that          : 785
be            : 777
which          : 722
;            : 633
or            : 581
from           : 551
but           : 533


-------------------------------------------------------
---------------------- DOCUMENT4 -----------------------
-------------------------------------------------------

Document contains a total of 83138  terms.
t            :3174
.            :2793
o            :2358
a            :2304
t            :2239
h            :1562
a            :1535
i            :1281
i            :1265
;            :1172
w            :1112
i            :1105
s            :1097
n            :1041
y            : 918
t            : 805
b            : 795
f            : 726
h            : 703
a            : 684
the           :3174
.            :2793
of            :2358
and            :2304
to            :2239
her           :1562
a            :1535
i            :1281
in            :1265
;            :1172
was            :1112
it            :1105
she           :1097
not           :1041
you            : 918
that          : 805
be            : 795
for           : 726
```

```
had          : 703
as           : 684


-----------------------------------------------------
--------------------- DOCUMENT5 ----------------------
-----------------------------------------------------

Document contains a total of 76232  terms.
t            :5833
.            :2803
o            :2370
a            :2121
a            :2092
i            :1893
t            :1583
i            :1370
w            :1335
o            : 976
f            : 726
t            : 722
i            : 703
a            : 643
h            : 550
m            : 533
w            : 527
s            : 518
f            : 509
a            : 488
the          :5833
.            :2803
of           :2370
and          :2121
a            :2092
i            :1893
to           :1583
in           :1370
was          :1335
on           : 976
for          : 726
that         : 722
it           : 703
at           : 643
had          : 550
my           : 533
with         : 527
spray        : 518
from         : 509
as           : 488
```

Krit Gupta
UIN: 927001565

```
-----------------------------------------------------
--------------------- DOCUMENT6 ----------------------
-----------------------------------------------------

Document contains a total of 35073  terms.
t            :2241
.            :1763
a             :1235
i            :1204
o             :1152
a             : 812
t            : 691
w              : 552
i            : 537
m              : 437
t            : 433
i            : 418
h             : 355
m              : 281
a             : 261
a             : 238
f            : 217
w              : 215
t            : 200
b             : 185
the           :2241
.            :1763
and           :1235
i            :1204
of            :1152
a             : 812
to            : 691
was            : 552
in            : 537
my             : 437
that          : 433
it           : 418
had           : 355
me            : 281
as            : 261
at            : 238
for           : 217
with          : 215
time          : 200
but           : 185
```

Krit Gupta
UIN: 927001565

```
-------------------------------------------------------
--------------------- DOCUMENT7 ----------------------
-------------------------------------------------------

Document contains a total of 80223  terms.
.            :3832
t            :3794
a             :3124
a             :1877
t            :1727
o             :1466
i            :1309
h             :1251
w             :1180
t            :1022
i            :1006
i            : 955
y             : 882
h             : 820
t            : 814
'            : 804
w             : 648
!            : 646
;            : 643
n             : 623
.            :3832
the           :3794
and           :3124
a             :1877
to            :1727
of            :1466
it           :1309
he            :1251
was           :1180
that          :1022
i            :1006
in            : 955
you           : 882
his           : 820
tom           : 814
's           : 804
with          : 648
!            : 646
;            : 643
n't           : 623
```

Krit Gupta
UIN: 927001565

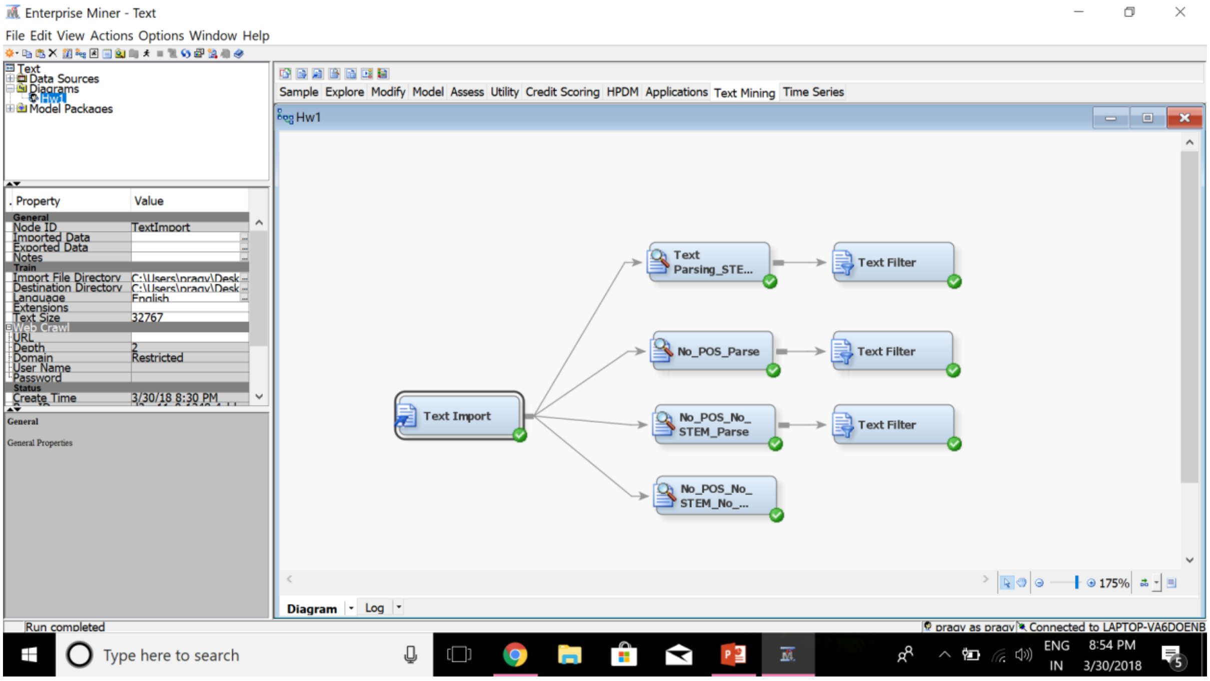
```
-------------------------------------------------------
--------------------- DOCUMENT8 ----------------------
-------------------------------------------------------

Document contains a total of 64515  terms.
t             :4794
.             :3004
a              :2503
o              :2301
a              :1635
i             :1295
t             :1175
i             :1001
w               : 854
t             : 793
i             : 688
h              : 582
m               : 472
a              : 451
w               : 449
a              : 444
h              : 427
o              : 378
w               : 369
f             : 348
the            :4794
.             :3004
and             :2503
of             :2301
a              :1635
i             :1295
to             :1175
in             :1001
was             : 854
that           : 793
it            : 688
had             : 582
my              : 472
as             : 451
with            : 449
at             : 444
he             : 427
on             : 378
were            : 369
for            : 348

Process finished with exit code 0
```

Krit Gupta
UIN: 927001565

SAS

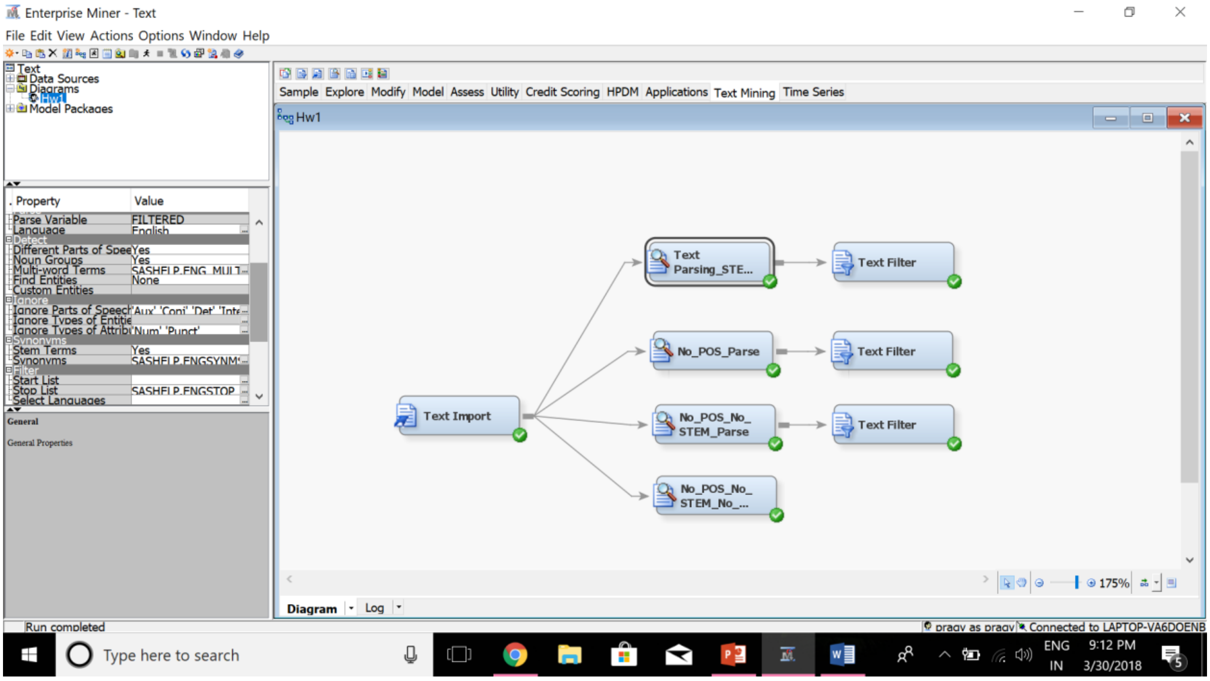


Property	Value
General	
Node ID	TextImport
Imported Data	
Exported Data	
Notes	
Train	
Import File Directory	C:\Users\pragy\Desk
Destination Directory	C:\Users\pragy\Desk
Language	English
Extensions	
Text Size	32767
Web Crawl	
URL	
Depth	2
Domain	Restricted
User Name	
Password	
Status	
Create Time	3/30/18 8:30 PM

General

General Properties

a. *Remove stopwords, apply POS and stemming.*



Property	Value
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Spee	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entitie	
Ignore Types of Attribu	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

General

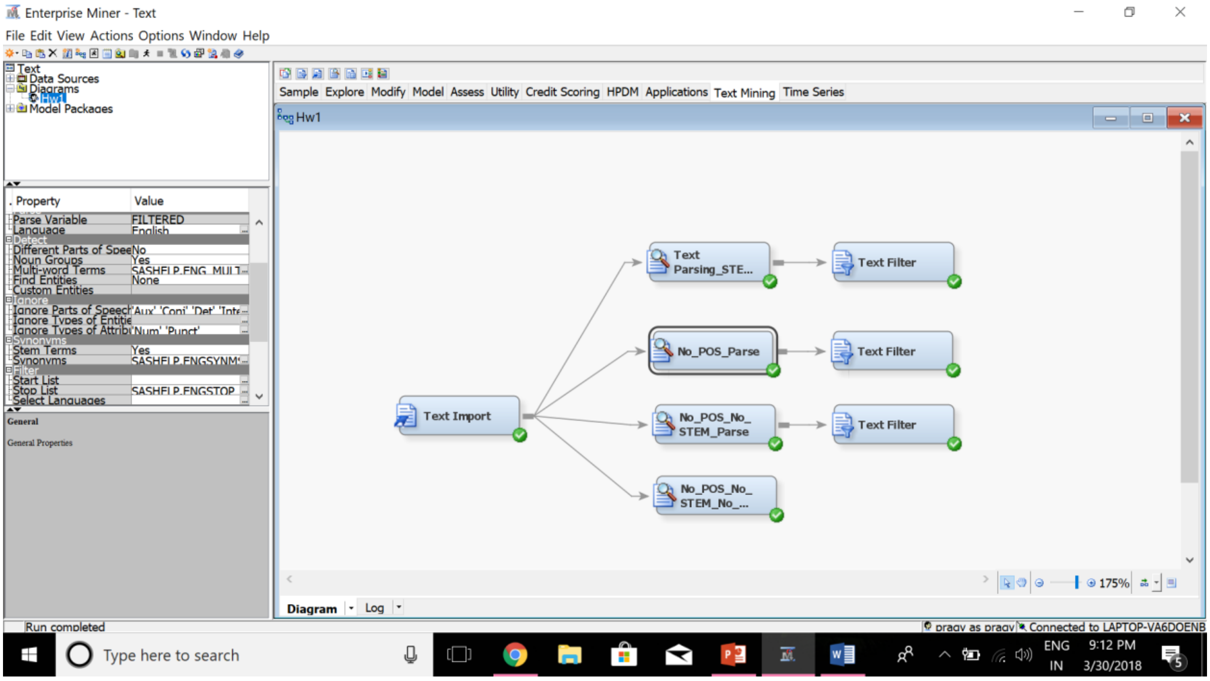
General Properties

```
+ water   Noun  Alpha  Keep  0.41954242214126  1533.0 1533.0 8.0   8.0 1.0       +  11482.0
+ man     Noun  Alpha  Keep  0.1119506873085   1112.0 1112.0 8.0   8.0 1.0       +  38024.0
+ time    Noun  Alpha  Keep  0.0096191071733   1017.0 1017.0 8.0   8.0 1.0       +  21043.0
+ know    Verb  Alpha  Keep  0.0516002953134   951.0   951.0  8.0   8.0 1.0       +  41532.0
+ little  Adj   Alpha  Keep  0.0213501206189   879.0   879.0  8.0   8.0 1.0       +  19569.0
```

Krit Gupta
UIN: 927001565

+ air	Noun	Alpha	Keep	0.3349677064352	862.0	862.0	8.0	8.0	1.0	+	387.0
+ day	Noun	Alpha	Keep	0.1137121474133	844.0	844.0	8.0	8.0	1.0	+	17927.0
+ thing	Noun	Alpha	Keep	0.0910603614383	824.0	824.0	8.0	8.0	1.0	+	22721.0
tom	Prop	Alpha	Keep	0.9827882834738	823.0	823.0	4.0	4.0	3118.0		14620.0
+ light	Noun	Alpha	Keep	0.2300646563978	774.0	774.0	8.0	8.0	1.0	+	12685.0
+ find	Verb	Alpha	Keep	0.0203769496465	722.0	722.0	8.0	8.0	1.0	+	31656.0
+ look	Verb	Alpha	Keep	0.0541680326657	651.0	651.0	8.0	8.0	1.0	+	30479.0
+ great	Adj	Alpha	Keep	0.0533400114462	621.0	621.0	8.0	8.0	1.0	+	7572.0
+ good	Adj	Alpha	Keep	0.0990676543699	587.0	587.0	8.0	8.0	1.0	+	3346.0
+ hand	Noun	Alpha	Keep	0.0439546920348	578.0	578.0	8.0	8.0	1.0	+	31943.0
+ fig	Noun	Alpha	Keep	0.6772355825861	575.0	575.0	3.0	3.0	4427.0	+	35250.0
+ eye	Noun	Alpha	Keep	0.0989578081226	558.0	558.0	8.0	8.0	1.0	+	5544.0
+ illustration	Noun	Alpha	Keep	0.4673151445266	544.0	544.0	6.0	6.0	1442.0	+	21708.0
+ turn	Verb	Alpha	Keep	0.0899168246168	525.0	525.0	8.0	8.0	1.0	+	5811.0
+ small	Adj	Alpha	Keep	0.1029459848069	486.0	486.0	8.0	8.0	1.0	+	21983.0

b. Remove stopwords, do not apply POS, apply stemming



Property	Value
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Spee	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entitie	
Ignore Types of Attribu	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

General

General Properties

Run completed

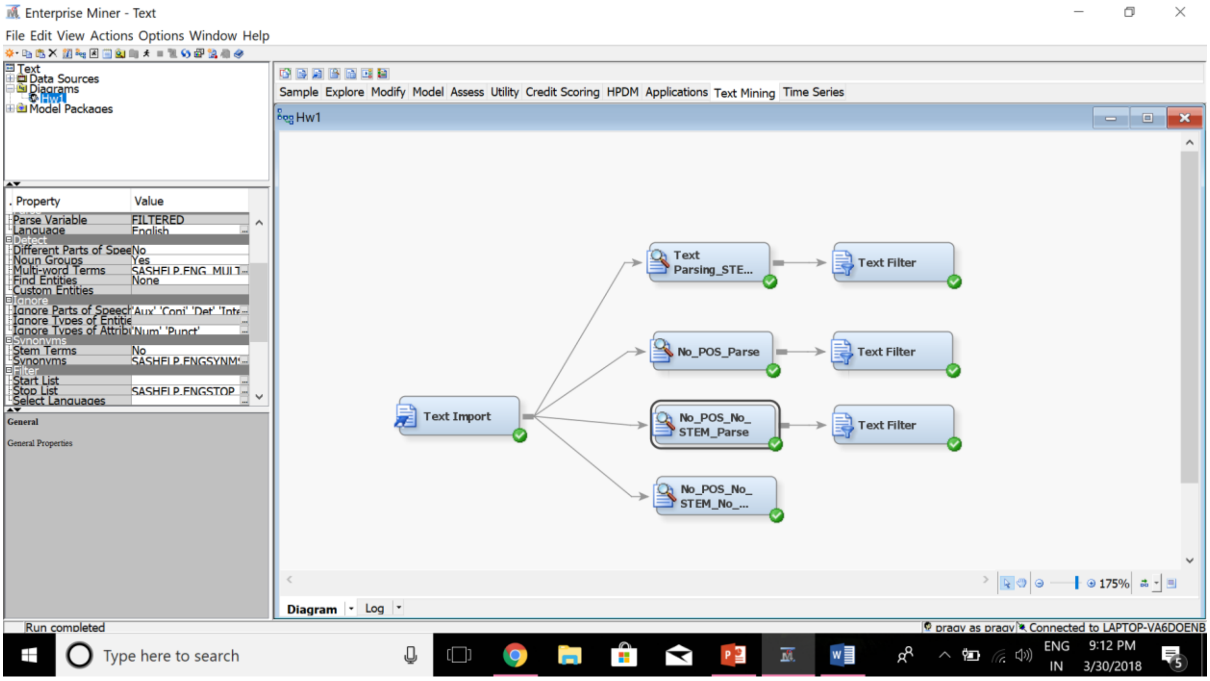
pragy as pragy Connected to LAPTOP-VA6DOENB

+ water	Alpha	Keep	0.43160396669823764	1983.0	1983.0	8.0	8.0	1.0	+ 1200.0
+ man	Alpha	Keep	0.09974608878241575	1246.0	1246.0	8.0	8.0	1.0	+ 33209.0
+ time	Alpha	Keep	0.010691289482238187	1223.0	1223.0	8.0	8.0	1.0	+ 3268.0
+ light	Alpha	Keep	0.20147575040479404	1177.0	1177.0	8.0	8.0	1.0	+ 34640.0
+ little	Alpha	Keep	0.009475754313934992	1124.0	1124.0	8.0	8.0	1.0	+ 26861.0

Krit Gupta
UIN: 927001565

```
+ air     Alpha Keep 0.37216585827855586   1096.0 1096.0 8.0 8.0 1.0      + 37269.0
+ know    Alpha Keep 0.04866045775792549   1036.0 1036.0 8.0 8.0 1.0      + 35205.0
+ day     Alpha Keep 0.11371214741339031   844.0  844.0  8.0 8.0 1.0      + 602.0
+ thing   Alpha Keep 0.089720683367283     829.0  829.0  8.0 8.0 1.0      + 24511.0
tom       Alpha Keep 0.9827882834738713    823.0  823.0  4.0 4.0 2636.0   632.0
+ good    Alpha Keep 0.08890933849726812   785.0  785.0  8.0 8.0 1.0      + 9185.0
+ find    Alpha Keep 0.021877132285045064 774.0  774.0  8.0 8.0 1.0      + 16629.0
+ look    Alpha Keep 0.059833785766625125 761.0  761.0  8.0 8.0 1.0      + 7782.0
+ well    Alpha Keep 0.07021659977115702   747.0  747.0  8.0 8.0 1.0      + 9339.0
+ great   Alpha Keep 0.0501673089652781    720.0  720.0  8.0 8.0 1.0      + 23832.0
+ heat    Alpha Keep 0.5788107399985035    709.0  709.0  6.0 6.0 1260.0 + 675.0
+ hand    Alpha Keep 0.03617151571695998   665.0  665.0  8.0 8.0 1.0      + 1753.0
+ place   Alpha Keep 0.042760884006016076 665.0  665.0  8.0 8.0 1.0      + 37827.0
+ long    Alpha Keep 0.024895780585701188 656.0  656.0  8.0 8.0 1.0      + 6418.0
+ turn    Alpha Keep 0.07287584737920016   635.0  635.0  8.0 8.0 1.0      + 15951.0
+ crowd   Alpha Keep 0.23337236740290557   108.0  108.0  8.0 8.0 1.0      + 1434.0
```

c. Remove stopwords, do not apply POS, do not stem



Property	Value
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Spee	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entitie	
Ignore Types of Attribu	'Num' 'Punct'
Synonyms	
Stem Terms	No
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

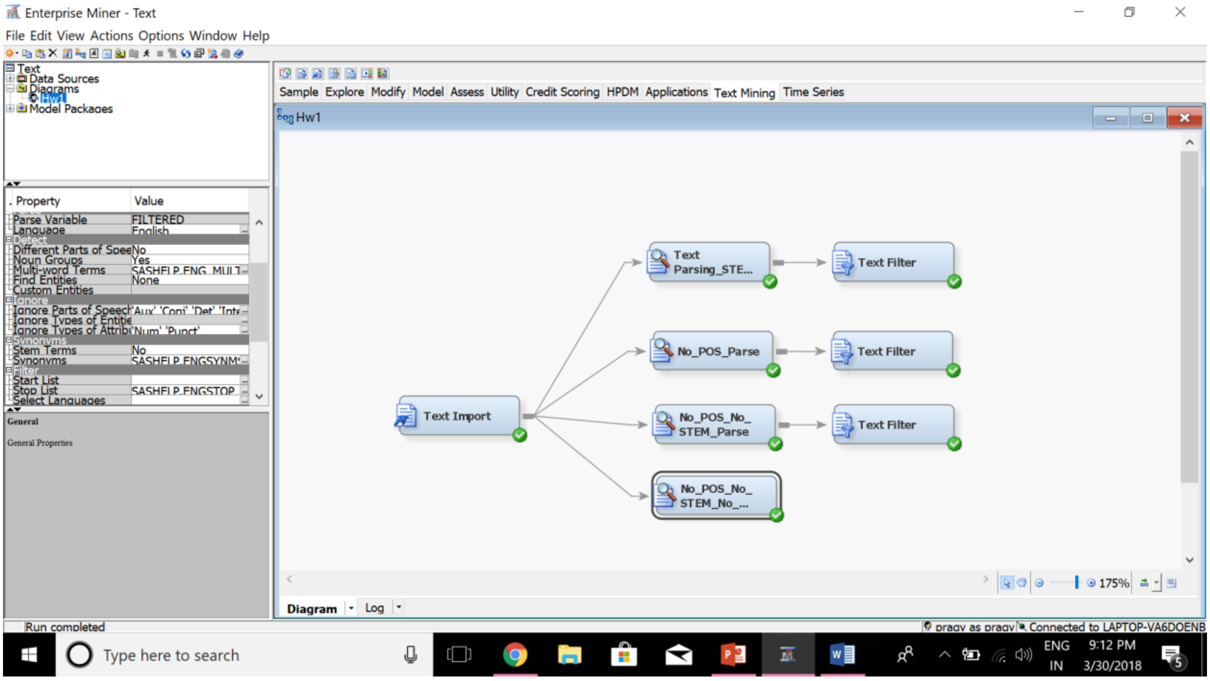
General

General Properties

```
water     Alpha Keep 0.44779067693756713   1943.0 1943.0 8.0 8.0 1.0      1377.0
air       Alpha Keep 0.3730874035877434    1093.0 1093.0 8.0 8.0 1.0      44716.0
time      Alpha Keep 0.017060069770459374 1057.0 1057.0 8.0 8.0 1.0      3801.0
light     Alpha Keep 0.25297631848502666   946.0  946.0  8.0 8.0 1.0      41590.0
tom       Alpha Keep 0.9827882834738713    823.0  823.0  4.0 4.0 3172.0   691.0
little    Alpha Keep 0.03119821550924784   802.0  802.0  8.0 8.0 1.0      32138.0
man       Alpha Keep 0.07172679820281158   799.0  799.0  8.0 8.0 1.0      39774.0
```

back	Alpha	Keep	0.07244425284634204	613.0	613.0	8.0	8.0	1.0	24504.0
know	Alpha	Keep	0.06639349493074942	599.0	599.0	8.0	8.0	1.0	42247.0
first	Alpha	Keep	0.031071107426880906	592.0	592.0	8.0	8.0	1.0	20071.0
fig	Alpha	Keep	0.6829486284433814	572.0	572.0	2.0	2.0	6906.0	20845.0
well	Alpha	Keep	0.11040126287247487	562.0	562.0	8.0	8.0	1.0	11076.0
day	Alpha	Keep	0.10138192354718045	558.0	558.0	8.0	8.0	1.0	652.0
heat	Alpha	Keep	0.5592977744958445	557.0	557.0	6.0	6.0	1387.0	734.0
great	Alpha	Keep	0.04023054192576403	553.0	553.0	8.0	8.0	1.0	28468.0
saw	Alpha	Keep	0.10068272995581906	545.0	545.0	8.0	8.0	1.0	21229.0
good	Alpha	Keep	0.11008576742586196	541.0	541.0	8.0	8.0	1.0	10895.0
illustration	Alpha	Keep	0.493768117435762	521.0	521.0	5.0	5.0	2135.0	31086.0
down	Alpha	Keep	0.08895935807198607	502.0	502.0	8.0	8.0	1.0	42898.0
people	Alpha	Keep	0.09086611123419153	492.0	492.0	8.0	8.0	1.0	45020.0

d. Do no remove stopwords, do not apply POS and do not stem.



Property	Value
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Spee	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entitie	
Ignore Types of Attrib	'Num' 'Punct'
Synonyms	
Stem Terms	No
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

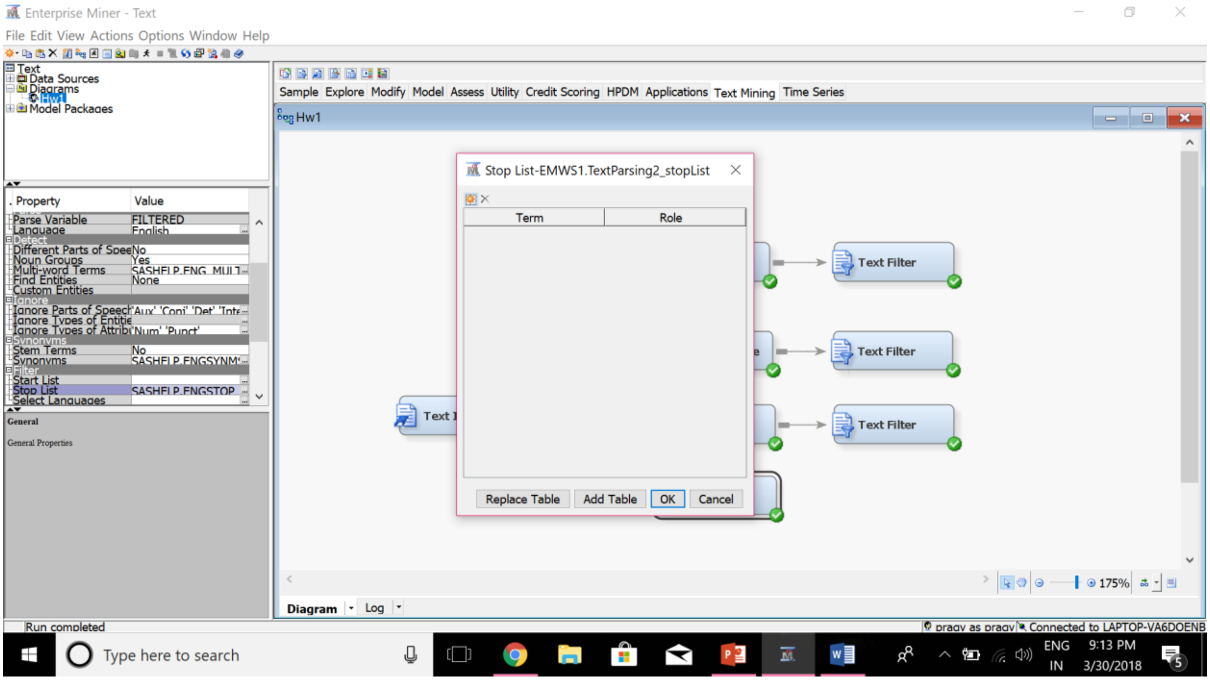
General

General Properties

Run completed

pragy as pragy Connected to LAPTOP-VA6DOENB

Krit Gupta
UIN: 927001565



Property	Value
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Spee	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULT
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Inte
Ignore Types of Entitie	
Ignore Types of Attribu	'Num' 'Punct'
Synonyms	
Stem Terms	No
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

was	Alpha	4445.0	8.0	Y	38867.0	1.0
not	Alpha	4204.0	8.0	Y	20397.0	1.0
is	Alpha	4025.0	8.0	Y	27761.0	1.0
s	Alpha	2238.0	8.0	Y	5182.0	1.0
be	Alpha	2148.0	8.0	Y	16454.0	1.0
water	Alpha	1943.0	8.0	Y	1394.0	1.0
have	Alpha	1912.0	8.0	Y	14343.0	1.0
no	Alpha	1655.0	8.0	Y	35209.0	1.0
then	Alpha	1361.0	8.0	Y	33565.0	1.0
do	Alpha	1303.0	8.0	Y	19971.0	1.0
said	Alpha	1272.0	8.0	Y	45191.0	1.0
were	Alpha	1225.0	8.0	Y	20160.0	1.0
one	Alpha	1214.0	8.0	Y	22617.0	1.0
now	Alpha	1115.0	8.0	Y	20517.0	1.0
more	Alpha	1100.0	8.0	Y	22571.0	1.0
air	Alpha	1093.0	8.0	Y	45083.0	1.0
very	Alpha	1073.0	8.0	Y	12147.0	1.0
time	Alpha	1057.0	8.0	Y	3829.0	1.0
so	Alpha	1014.0	8.0	Y	42799.0	1.0
had	Alpha	997.0	8.0	Y	17544.0	1.0