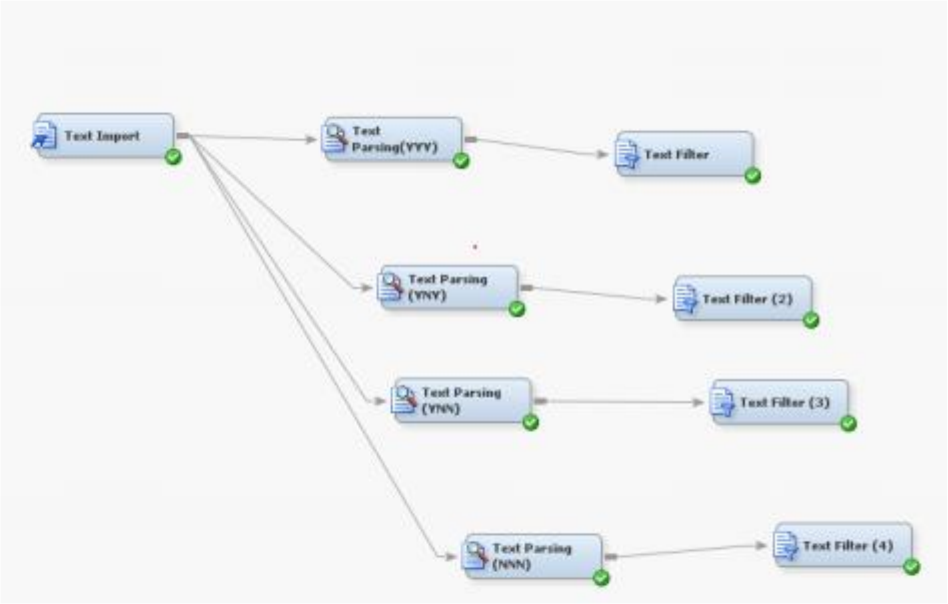
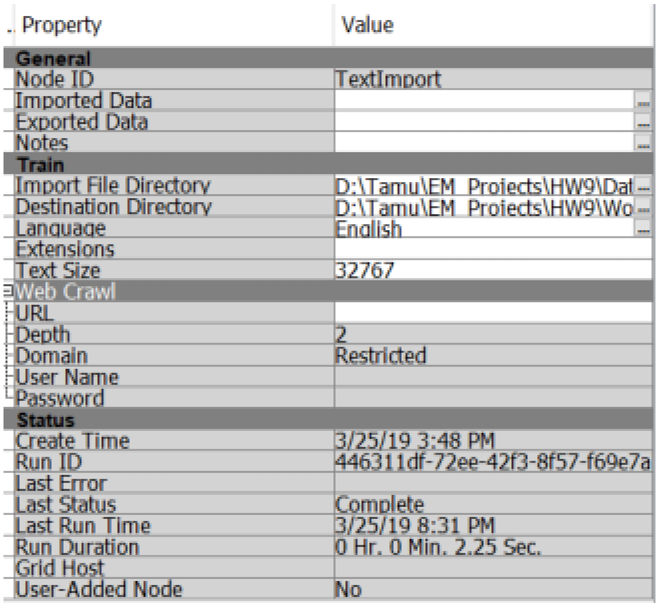
SAS Part

Q1

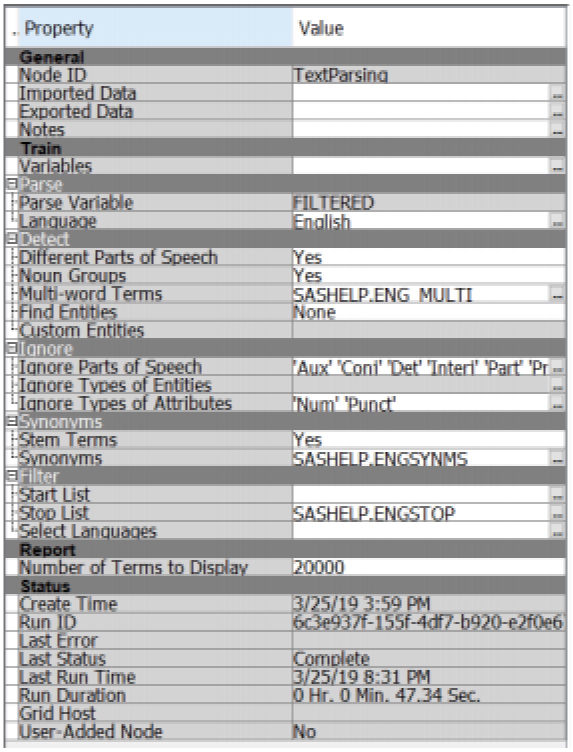


2. Import Property



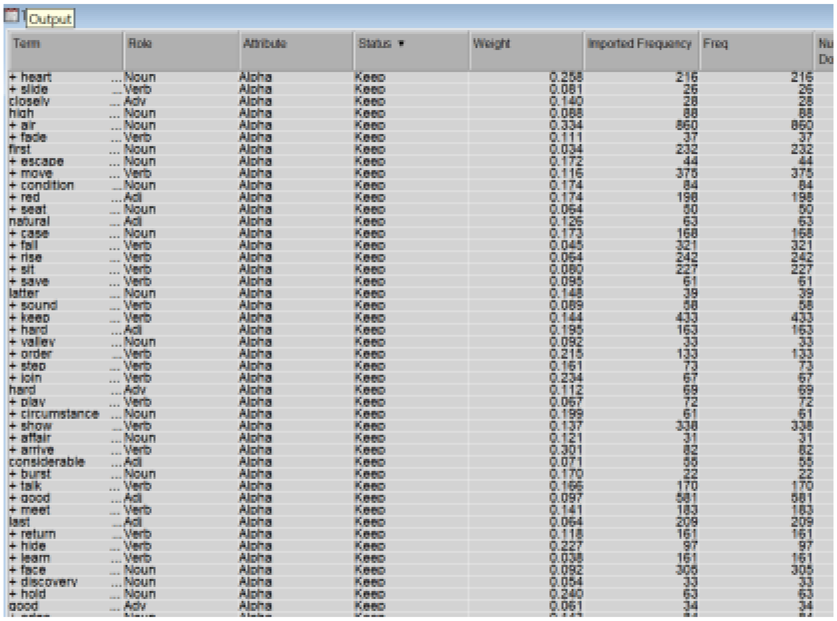
Property	Value
General	
Node ID	TextImport
Imported Data	
Exported Data	
Notes	
Train	
Import File Directory	D:\Tamu\EM_Projects\HW9\Dat
Destination Directory	D:\Tamu\EM_Projects\HW9\Wo
Language	English
Extensions	
Text Size	32767
Web Crawl	
URL	
Depth	2
Domain	Restricted
User Name	
Password	
Status	
Create Time	3/25/19 3:48 PM
Run ID	446311df-72ee-42f3-8f57-f69e7a
Last Error	
Last Status	Complete
Last Run Time	3/25/19 8:31 PM
Run Duration	0 Hr. 0 Min. 2.25 Sec.
Grid Host	
User-Added Node	No

a. Stop Words-Yes; POS-Yes; Stem-Yes



Property	Value
General	
Node ID	TextParsing
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Parse	
Parse Variable	FILTERED
Language	English ...
⊟Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI ...
Find Entities	None
Custom Entities	
⊟Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Pr ...
Ignore Types of Entities	...
Ignore Types of Attributes	'Num' 'Punct' ...
⊟Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS ...
⊟Filter	
Start List	...
Stop List	SASHELP.ENGSTOP ...
Select Languages	...
Report	
Number of Terms to Display	20000
Status	
Create Time	3/25/19 3:59 PM
Run ID	6c3e937f-155f-4df7-b920-e2f0e6
Last Error	
Last Status	Complete
Last Run Time	3/25/19 8:31 PM
Run Duration	0 Hr. 0 Min. 47.34 Sec.
Grid Host	
User-Added Node	No

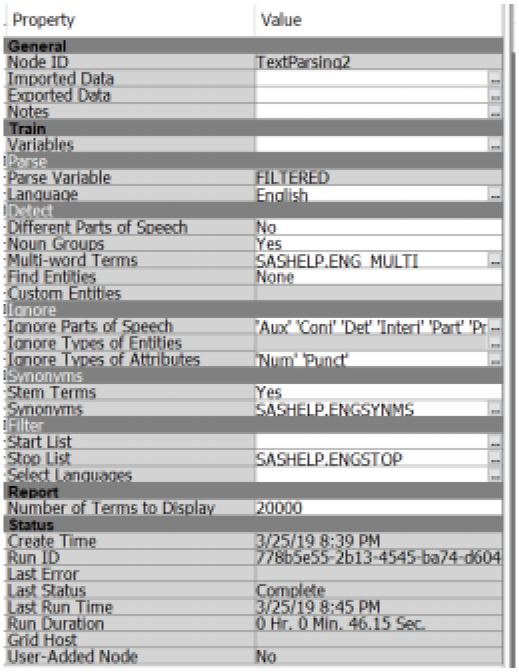
Number of terms = 2500



Output

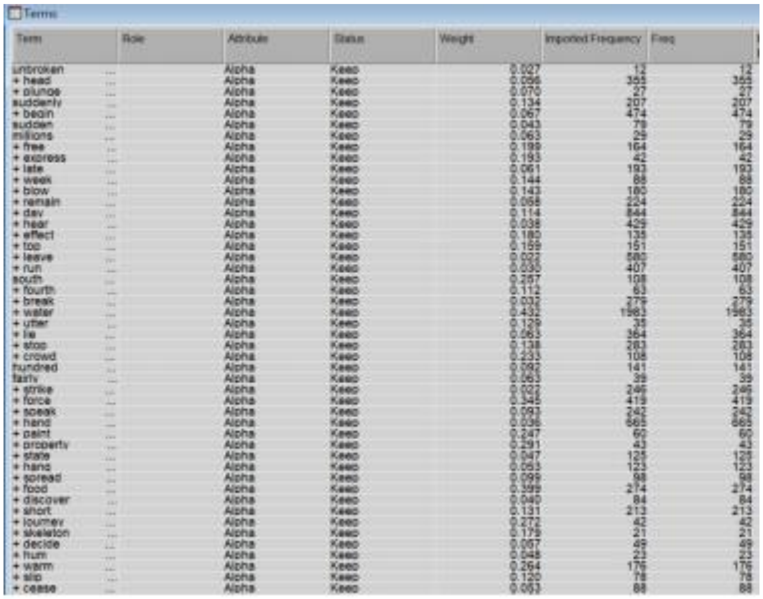
Term	Role	Attribute	Status ▾	Weight	Imported Frequency	Freq	Nu Do
+ heart ...	Noun	Alpha	Keep	0.258	216	216	
+ slide ...	Verb	Alpha	Keep	0.081	26	26	
closely ...	Adv	Alpha	Keep	0.140	28	28	
high ...	Noun	Alpha	Keep	0.088	88	88	
+ air ...	Noun	Alpha	Keep	0.334	860	860	
+ fade ...	Verb	Alpha	Keep	0.111	37	37	
first ...	Noun	Alpha	Keep	0.034	232	232	
+ escape ...	Noun	Alpha	Keep	0.172	44	44	
+ move ...	Verb	Alpha	Keep	0.116	375	375	
+ condition ...	Noun	Alpha	Keep	0.174	84	84	
+ red ...	Adj	Alpha	Keep	0.174	198	198	
+ seat ...	Noun	Alpha	Keep	0.064	50	50	
natural ...	Adj	Alpha	Keep	0.126	63	63	
+ case ...	Noun	Alpha	Keep	0.173	168	168	
+ fall ...	Verb	Alpha	Keep	0.045	321	321	
+ rise ...	Verb	Alpha	Keep	0.064	242	242	
+ sit ...	Verb	Alpha	Keep	0.080	227	227	
+ save ...	Verb	Alpha	Keep	0.095	61	61	
latter ...	Noun	Alpha	Keep	0.148	39	39	
+ sound ...	Verb	Alpha	Keep	0.089	58	58	
+ keep ...	Verb	Alpha	Keep	0.144	433	433	
+ hard ...	Adj	Alpha	Keep	0.195	163	163	
+ valley ...	Noun	Alpha	Keep	0.092	33	33	
+ order ...	Verb	Alpha	Keep	0.215	133	133	
+ step ...	Verb	Alpha	Keep	0.161	73	73	
+ join ...	Verb	Alpha	Keep	0.234	67	67	
hard ...	Adv	Alpha	Keep	0.112	69	69	
+ play ...	Verb	Alpha	Keep	0.067	72	72	
+ circumstance ...	Noun	Alpha	Keep	0.199	61	61	
+ show ...	Verb	Alpha	Keep	0.137	338	338	
+ affair ...	Noun	Alpha	Keep	0.121	31	31	
+ arrive ...	Verb	Alpha	Keep	0.301	82	82	
considerable ...	Adj	Alpha	Keep	0.071	55	55	
+ burst ...	Noun	Alpha	Keep	0.170	22	22	
+ talk ...	Verb	Alpha	Keep	0.166	170	170	
+ good ...	Adj	Alpha	Keep	0.097	581	581	
+ meet ...	Verb	Alpha	Keep	0.141	183	183	
last ...	Adj	Alpha	Keep	0.064	209	209	
+ return ...	Verb	Alpha	Keep	0.118	161	161	
+ hide ...	Verb	Alpha	Keep	0.227	97	97	
+ learn ...	Verb	Alpha	Keep	0.038	161	161	
+ face ...	Noun	Alpha	Keep	0.092	305	305	
+ discovery ...	Noun	Alpha	Keep	0.054	33	33	
+ hold ...	Noun	Alpha	Keep	0.240	63	63	
good ...	Adv	Alpha	Keep	0.061	34	34	

b) Stop Words-Yes; POS-No; Stem-Yes



Property	Value
General	
Node ID	TextParsing2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Parse	
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Speech	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Pr
Ignore Types of Entities	
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	
Report	
Number of Terms to Display	20000
Status	
Create Time	3/25/19 8:39 PM
Run ID	778b5e55-2b13-4545-ba74-d604
Last Error	
Last Status	Complete
Last Run Time	3/25/19 8:45 PM
Run Duration	0 Hr. 0 Min. 46.15 Sec.
Grid Host	
User-Added Node	No

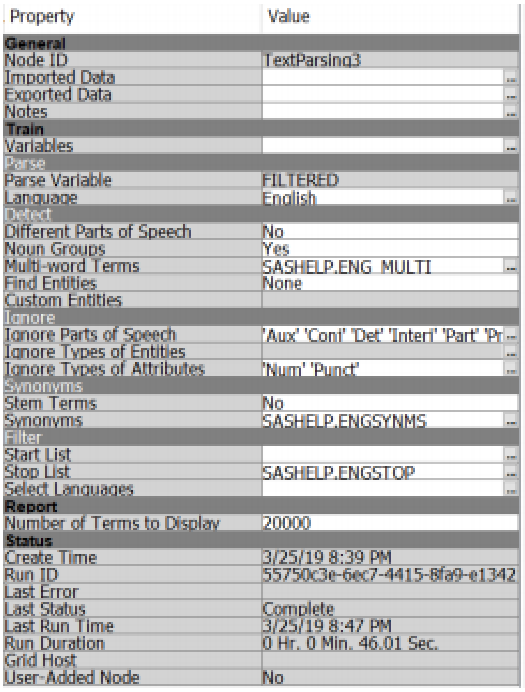
Number of terms=4944



Terms

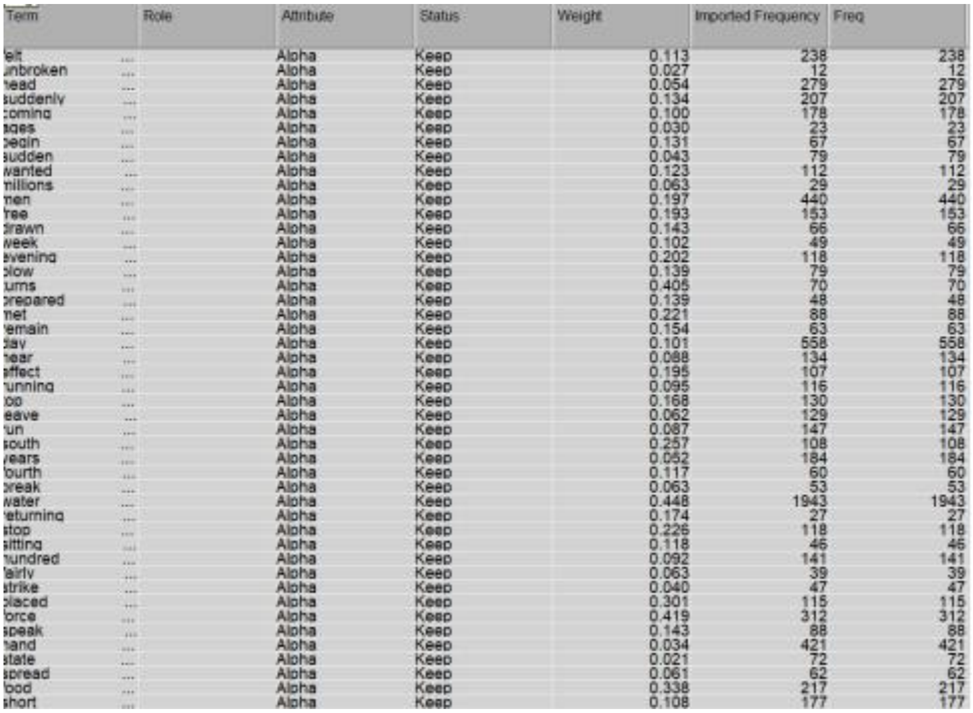
Term	Role	Attribute	Status	Weight	Imported Frequency	Freq
unbroken		Alpha	Keep	0.027	12	12
+ head		Alpha	Keep	0.056	355	355
+ plunge		Alpha	Keep	0.070	27	27
suddenly		Alpha	Keep	0.134	207	207
+ begin		Alpha	Keep	0.067	474	474
sudden		Alpha	Keep	0.043	79	79
millions		Alpha	Keep	0.063	29	29
+ free		Alpha	Keep	0.199	164	164
+ express		Alpha	Keep	0.193	42	42
+ late		Alpha	Keep	0.061	193	193
+ week		Alpha	Keep	0.144	88	88
+ blow		Alpha	Keep	0.143	180	180
+ remain		Alpha	Keep	0.068	224	224
+ day		Alpha	Keep	0.114	844	844
+ hear		Alpha	Keep	0.038	429	429
+ effect		Alpha	Keep	0.180	135	135
+ top		Alpha	Keep	0.159	151	151
+ leave		Alpha	Keep	0.022	580	580
+ run		Alpha	Keep	0.030	407	407
south		Alpha	Keep	0.267	108	108
+ fourth		Alpha	Keep	0.112	63	63
+ break		Alpha	Keep	0.032	279	279
+ water		Alpha	Keep	0.432	1983	1983
+ utter		Alpha	Keep	0.129	35	35
+ lie		Alpha	Keep	0.063	364	364
+ stop		Alpha	Keep	0.138	283	283
+ crowd		Alpha	Keep	0.233	108	108
hundred		Alpha	Keep	0.092	141	141
fairly		Alpha	Keep	0.063	39	39
+ strike		Alpha	Keep	0.022	246	246
+ force		Alpha	Keep	0.345	419	419
+ speak		Alpha	Keep	0.093	242	242
+ hand		Alpha	Keep	0.036	665	665
+ paint		Alpha	Keep	0.247	60	60
+ property		Alpha	Keep	0.291	43	43
+ state		Alpha	Keep	0.047	125	125
+ hang		Alpha	Keep	0.053	123	123
+ spread		Alpha	Keep	0.099	98	98
+ food		Alpha	Keep	0.399	274	274
+ discover		Alpha	Keep	0.040	84	84
+ short		Alpha	Keep	0.131	213	213
+ journey		Alpha	Keep	0.272	42	42
+ skeleton		Alpha	Keep	0.179	21	21
+ decide		Alpha	Keep	0.067	49	49
+ hum		Alpha	Keep	0.048	23	23
+ warm		Alpha	Keep	0.264	176	176
+ slip		Alpha	Keep	0.120	78	78
+ cease		Alpha	Keep	0.053	88	88

c) Stop Words-Yes; POS-No; Stem-No



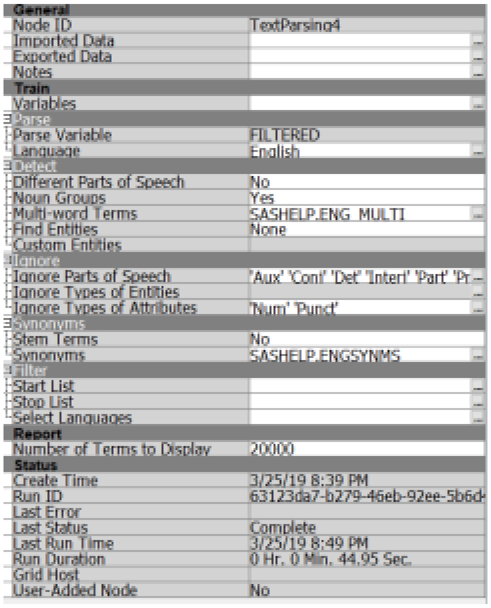
Property	Value
General	
Node ID	TextParsing3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Parse	
Parse Variable	FILTERED
Language	English ...
Detect	
Different Parts of Speech	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI ...
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Pr ...
Ignore Types of Entities	...
Ignore Types of Attributes	'Num' 'Punct' ...
Synonyms	
Stem Terms	No
Synonyms	SASHELP.ENGSYNMS ...
Filter	
Start List	...
Stop List	SASHELP.ENGSTOP ...
Select Languages	...
Report	
Number of Terms to Display	20000
Status	
Create Time	3/25/19 8:39 PM
Run ID	55750c3e-6ec7-4415-8fa9-e1342
Last Error	
Last Status	Complete
Last Run Time	3/25/19 8:47 PM
Run Duration	0 Hr. 0 Min. 46.01 Sec.
Grid Host	
User-Added Node	No

Number of terms=5024



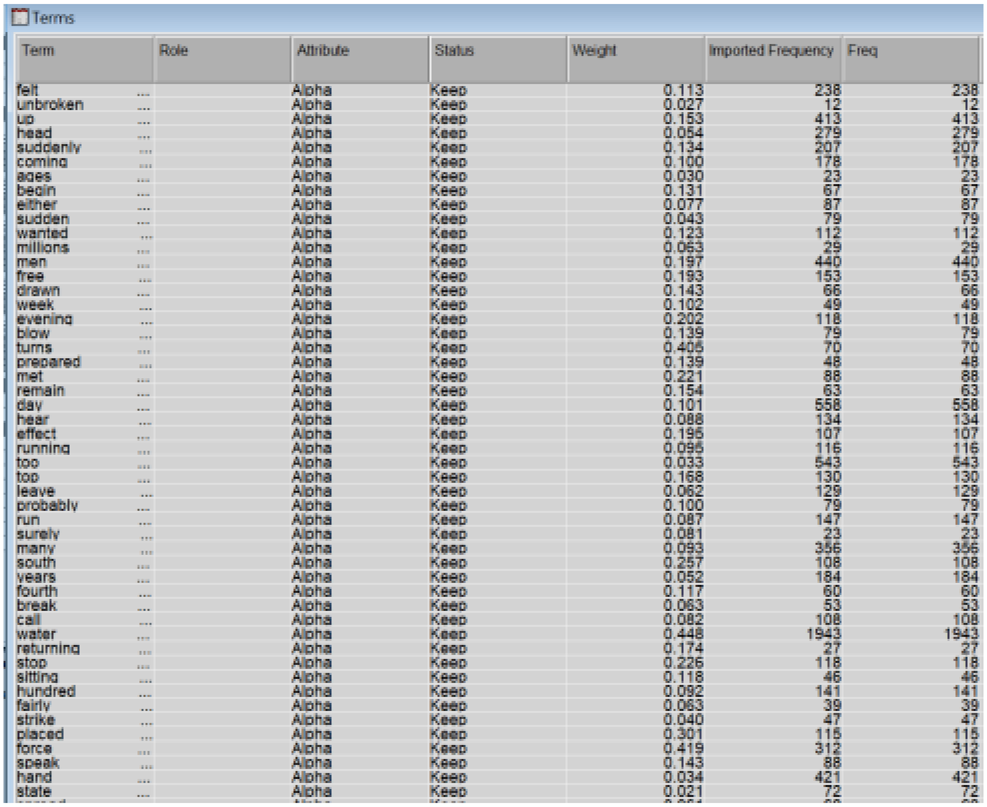
Term	Role	Attribute	Status	Weight	Imported Frequency	Freq
felt ...		Alpha	Keep	0.113	238	238
unbroken ...		Alpha	Keep	0.027	12	12
head ...		Alpha	Keep	0.054	279	279
suddenly ...		Alpha	Keep	0.134	207	207
coming ...		Alpha	Keep	0.100	178	178
ages ...		Alpha	Keep	0.030	23	23
begin ...		Alpha	Keep	0.131	67	67
sudden ...		Alpha	Keep	0.043	79	79
wanted ...		Alpha	Keep	0.123	112	112
millions ...		Alpha	Keep	0.063	29	29
men ...		Alpha	Keep	0.197	440	440
free ...		Alpha	Keep	0.193	153	153
drawn ...		Alpha	Keep	0.143	66	66
week ...		Alpha	Keep	0.102	49	49
evening ...		Alpha	Keep	0.202	118	118
blow ...		Alpha	Keep	0.139	79	79
turns ...		Alpha	Keep	0.405	70	70
prepared ...		Alpha	Keep	0.139	48	48
met ...		Alpha	Keep	0.221	88	88
remain ...		Alpha	Keep	0.154	63	63
day ...		Alpha	Keep	0.101	558	558
hear ...		Alpha	Keep	0.088	134	134
affect ...		Alpha	Keep	0.195	107	107
running ...		Alpha	Keep	0.095	116	116
top ...		Alpha	Keep	0.168	130	130
leave ...		Alpha	Keep	0.062	129	129
run ...		Alpha	Keep	0.087	147	147
south ...		Alpha	Keep	0.257	108	108
years ...		Alpha	Keep	0.052	184	184
fourth ...		Alpha	Keep	0.117	60	60
break ...		Alpha	Keep	0.063	53	53
water ...		Alpha	Keep	0.448	1943	1943
returning ...		Alpha	Keep	0.174	27	27
stop ...		Alpha	Keep	0.226	118	118
sitting ...		Alpha	Keep	0.118	46	46
hundred ...		Alpha	Keep	0.092	141	141
fairly ...		Alpha	Keep	0.063	39	39
strike ...		Alpha	Keep	0.040	47	47
placed ...		Alpha	Keep	0.301	115	115
force ...		Alpha	Keep	0.419	312	312
speak ...		Alpha	Keep	0.143	88	88
hand ...		Alpha	Keep	0.034	421	421
state ...		Alpha	Keep	0.021	72	72
spread ...		Alpha	Keep	0.061	62	62
food ...		Alpha	Keep	0.338	217	217
short ...		Alpha	Keep	0.108	177	177

d) Stop Words-No; POS-No; Stem-No



General	
Node ID	TextParsing4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Parse	
Parse Variable	FILTERED
Language	English
Detect	
Different Parts of Speech	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Pr
Ignore Types of Entities	
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	No
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	
Select Languages	
Report	
Number of Terms to Display	20000
Status	
Create Time	3/25/19 8:39 PM
Run ID	63123da7-b279-46eb-92ee-5b6d
Last Error	
Last Status	Complete
Last Run Time	3/25/19 8:49 PM
Run Duration	0 Hr. 0 Min. 44.95 Sec.
Grid Host	
User-Added Node	No

Number of terms=6280



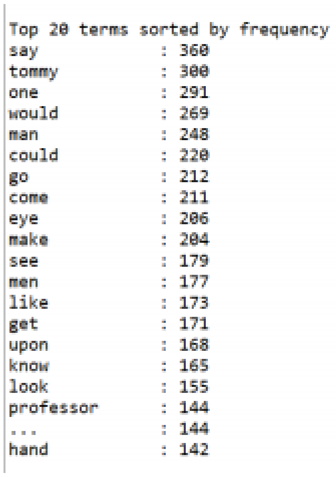
Terms

Term		Role	Attribute	Status	Weight	Imported Frequency	Freq
felt	...		Alpha	Keep	0.113	238	238
unbroken	...		Alpha	Keep	0.027	12	12
up	...		Alpha	Keep	0.153	413	413
head	...		Alpha	Keep	0.054	279	279
suddenly	...		Alpha	Keep	0.134	207	207
coming	...		Alpha	Keep	0.100	178	178
ages	...		Alpha	Keep	0.030	23	23
begin	...		Alpha	Keep	0.131	67	67
either	...		Alpha	Keep	0.077	87	87
sudden	...		Alpha	Keep	0.043	79	79
wanted	...		Alpha	Keep	0.123	112	112
millions	...		Alpha	Keep	0.063	29	29
men	...		Alpha	Keep	0.197	440	440
free	...		Alpha	Keep	0.193	153	153
drawn	...		Alpha	Keep	0.143	66	66
week	...		Alpha	Keep	0.102	49	49
evening	...		Alpha	Keep	0.202	118	118
blow	...		Alpha	Keep	0.139	79	79
turns	...		Alpha	Keep	0.405	70	70
prepared	...		Alpha	Keep	0.139	48	48
met	...		Alpha	Keep	0.221	88	88
remain	...		Alpha	Keep	0.154	63	63
day	...		Alpha	Keep	0.101	558	558
hear	...		Alpha	Keep	0.088	134	134
effect	...		Alpha	Keep	0.196	107	107
running	...		Alpha	Keep	0.096	116	116
too	...		Alpha	Keep	0.033	543	543
top	...		Alpha	Keep	0.168	130	130
leave	...		Alpha	Keep	0.062	129	129
probably	...		Alpha	Keep	0.100	79	79
run	...		Alpha	Keep	0.087	147	147
surely	...		Alpha	Keep	0.081	23	23
many	...		Alpha	Keep	0.093	356	356
south	...		Alpha	Keep	0.257	108	108
years	...		Alpha	Keep	0.052	184	184
fourth	...		Alpha	Keep	0.117	60	60
break	...		Alpha	Keep	0.063	53	53
call	...		Alpha	Keep	0.082	108	108
water	...		Alpha	Keep	0.448	1943	1943
returning	...		Alpha	Keep	0.174	27	27
stop	...		Alpha	Keep	0.226	118	118
sitting	...		Alpha	Keep	0.118	46	46
hundred	...		Alpha	Keep	0.092	141	141
fairly	...		Alpha	Keep	0.063	39	39
strike	...		Alpha	Keep	0.040	47	47
placed	...		Alpha	Keep	0.301	115	115
force	...		Alpha	Keep	0.419	312	312
speak	...		Alpha	Keep	0.143	88	88
hand	...		Alpha	Keep	0.034	421	421
state	...		Alpha	Keep	0.021	72	72

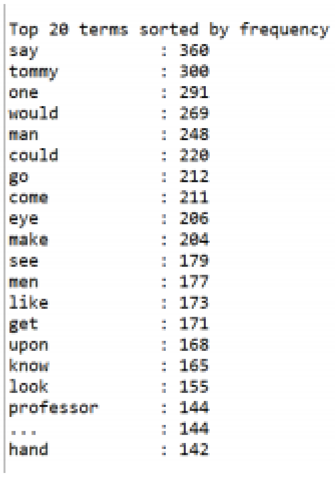
Python Part

```
import string
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import wordnet as wn
from nltk.corpus import stopwords
from nltk.probability import FreqDist
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('stopwords')
nltk.download('wordnet')
with open ("TextFiles/T1.txt", "r") as text_file:
adoc = text_file.read()

# Convert to all lower case - required
a_discussion = ("%s" %adoc).lower()

# Remove unwanted punctuation
a_discussion = a_discussion.replace
("''" != word) and ("``" != word) and (word!='description') and (word
!='dtype') \
and (word != 'object') and (word!="'s")]
print("\nDocument contains a total of", len(tokens), " terms.")

#POS Tagging
tagged_tokens = nltk.pos_tag(tokens)
pos_list = [word[1] for word in tagged_tokens if word[1] != ":" and \
word[1] != "."]
pos_dist = FreqDist(pos_list)
pos_dist.plot(title="Parts of Speech")
for pos, frequency in pos_dist.most_common(pos_dist.N()):
 print('{:<15s}:{:>4d}'.format(pos, frequency))

# 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 stop_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" ]
token_dist = FreqDist(stop_tokens)
print("\nCorpus contains", len(token_dist.items()), \
" unique terms after removing stop words.\n")
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))

#Token distribution
fdist = FreqDist(stemmed_tokens)
print("\nCorpus contains", len(fdist.items())," unique terms after
Stemming.\n")

# Top 20 terms
print('Top 20 terms sorted by frequency')
for word, freq in fdist.most_common(20):
 print('{:<15s}:{:>4d}'.format(word, freq))
fdist_top = nltk.probability.FreqDist()
for word, freq in fdist.most_common(20):
 fdist_top[word] = freq
fdist_top.plot()
```

Document 1 – Case 1

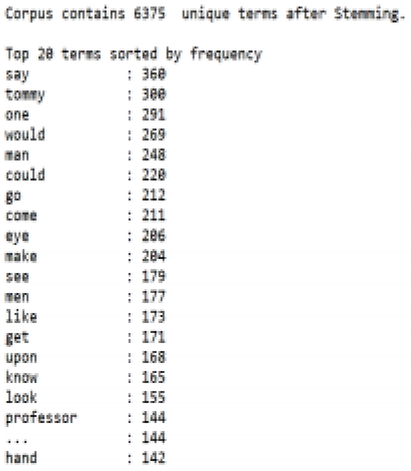


```
Top 20 terms sorted by frequency
say            : 360
tommy          : 300
one            : 291
would          : 269
man            : 248
could          : 220
go             : 212
come           : 211
eye            : 206
make           : 204
see            : 179
men            : 177
like           : 173
get            : 171
upon           : 168
know           : 165
look           : 155
professor      : 144
...            : 144
hand           : 142
```



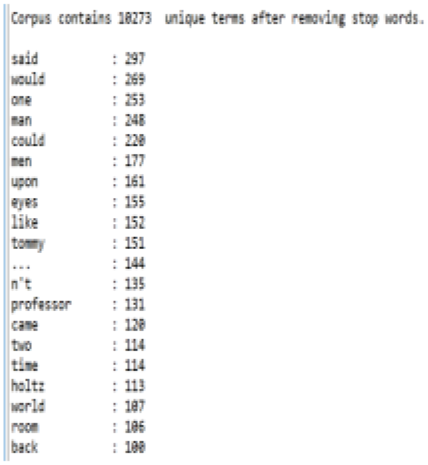
```
Top 20 terms sorted by frequency
say            : 360
tommy          : 300
one            : 291
would          : 269
man            : 248
could          : 220
go             : 212
come           : 211
eye            : 206
make           : 204
see            : 179
men            : 177
like           : 173
get            : 171
upon           : 168
know           : 165
look           : 155
professor      : 144
...            : 144
hand           : 142
```

Case 2



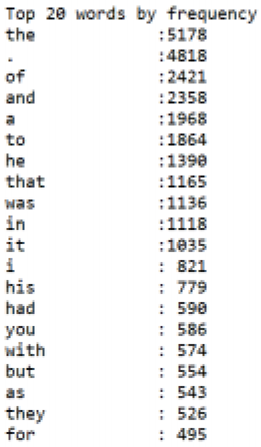
```
Corpus contains 6375  unique terms after Stemming.

Top 20 terms sorted by frequency
say             : 360
tommy           : 300
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
```

Case 3



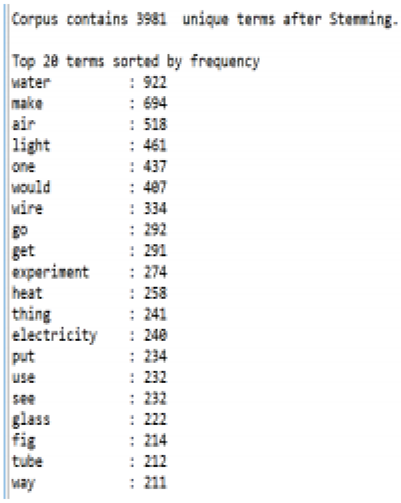
```
Corpus contains 10273  unique terms after removing stop words.

said            : 297
would           : 269
one             : 253
man             : 248
could           : 220
men             : 177
upon            : 161
eyes            : 155
like            : 152
tommy           : 151
...             : 144
n't             : 135
professor       : 131
came            : 120
two             : 114
time            : 114
holtz           : 113
world           : 107
room            : 106
back            : 100
```

Case 4

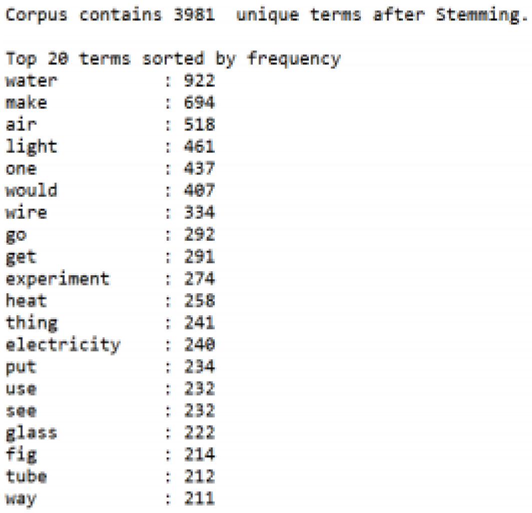


```
Top 20 words by frequency
the             :5178
.               :4818
of              :2421
and             :2358
a               :1968
to              :1864
he              :1390
that            :1165
was             :1136
in              :1118
it              :1035
i               : 821
his             : 779
had             : 590
you             : 586
with            : 574
but             : 554
as              : 543
they            : 526
for             : 495
```

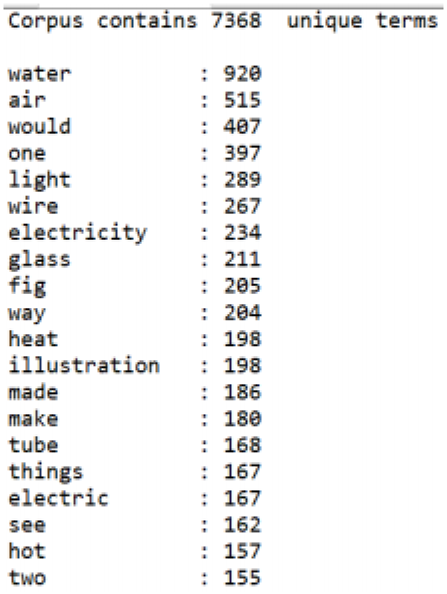
Document 2 – 4 cases in order



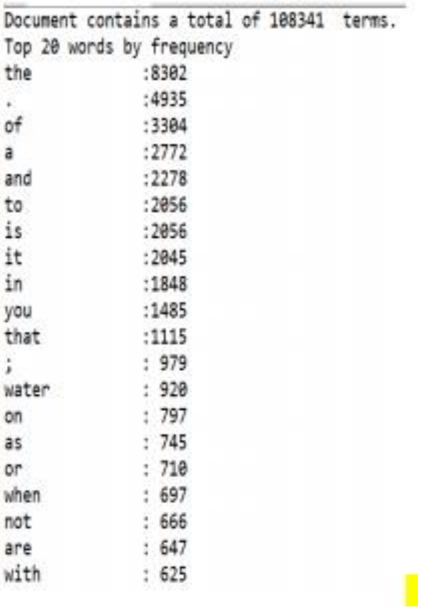
```
Corpus contains 3981  unique terms after Stemming.

Top 20 terms sorted by frequency
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
```



```
Corpus contains 3981  unique terms after Stemming.

Top 20 terms sorted by frequency
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
```

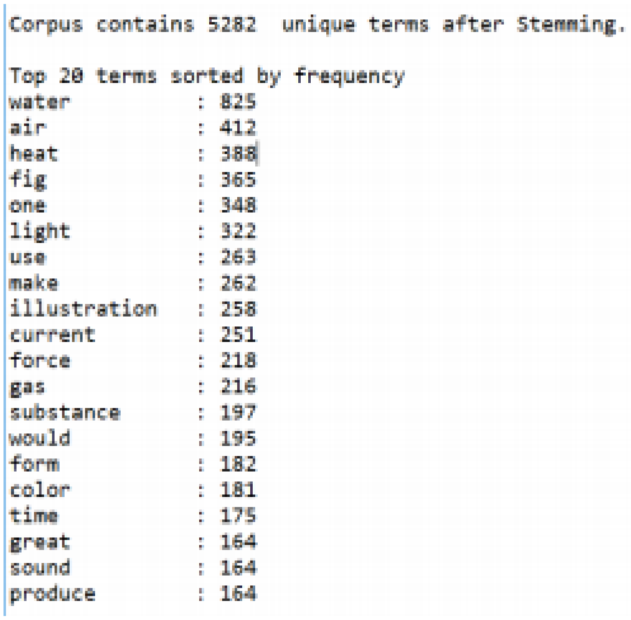


```
Corpus contains 7368  unique terms

water            : 920
air              : 515
would            : 407
one              : 397
light            : 289
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
```

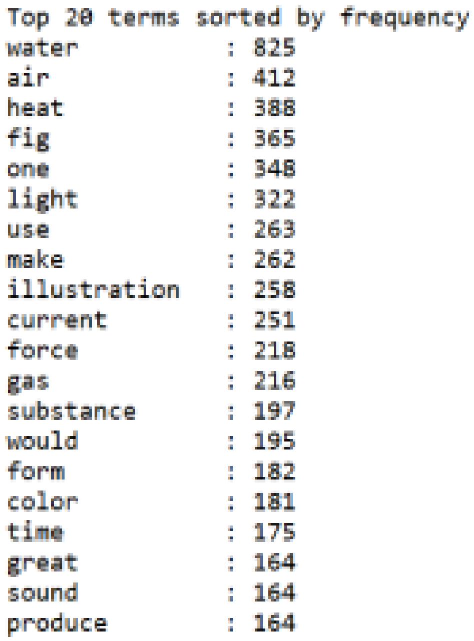


```
Document contains a total of 108341  terms.
Top 20 words by frequency
the             :8302
.               :4935
of              :3304
a               :2772
and             :2278
to              :2056
is              :2056
it              :2045
in              :1848
you             :1485
that            :1115
;               : 979
water           : 920
on              : 797
as              : 745
or              : 710
when            : 697
not             : 666
are             : 647
with            : 625
```

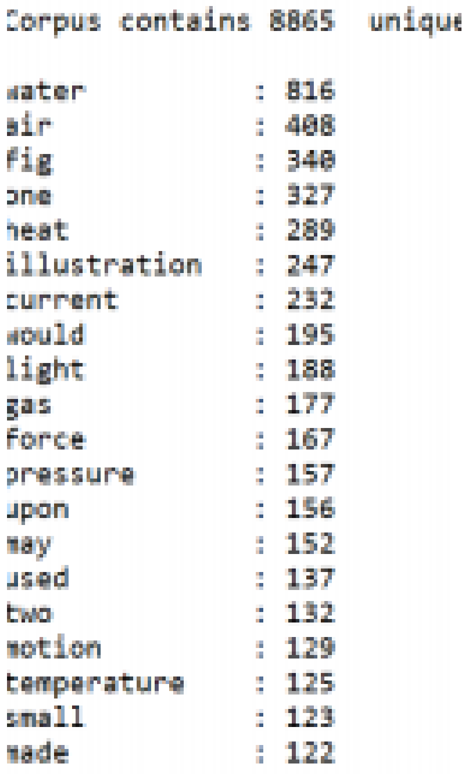
Document 3 – all 4 cases



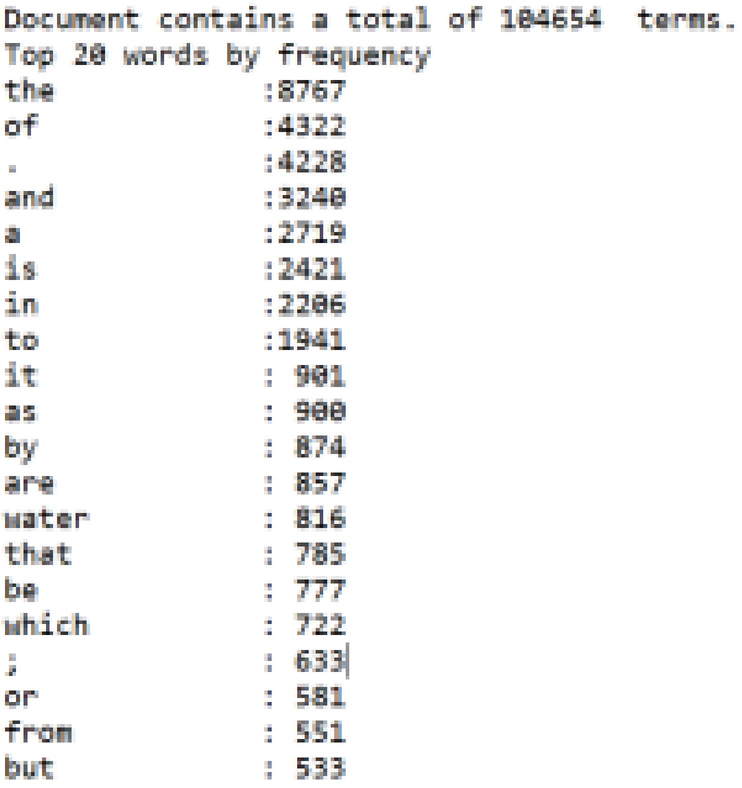
```
Corpus contains 5282  unique terms after Stemming.

Top 20 terms sorted by frequency
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
```



```
Top 20 terms sorted by frequency
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
```

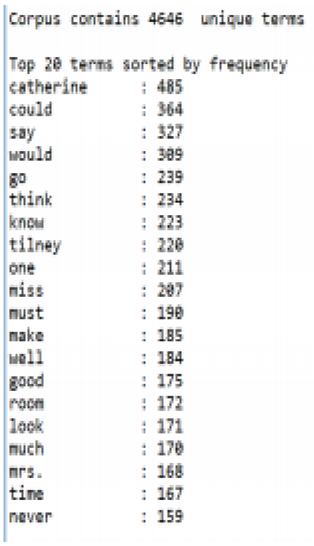


```
Corpus contains 8865   unique

water              : 816
air                : 408
fig                : 340
one                : 327
heat               : 289
illustration       : 247
current            : 232
would              : 195
light              : 188
gas                : 177
force              : 167
pressure           : 157
upon               : 156
may                : 152
used               : 137
two                : 132
motion             : 129
temperature        : 125
small              : 123
made               : 122
```

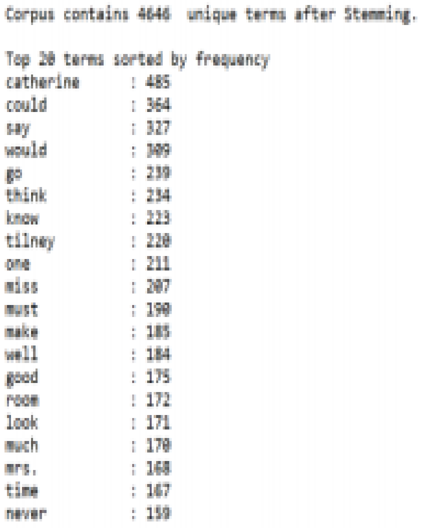


```
Document contains a total of 104654  terms.
Top 20 words by frequency
the                 :8767
of                  :4322
.                   :4228
and                 :3240
a                   :2719
is                  :2421
in                  :2206
to                  :1941
it                  : 901
as                  : 900
by                  : 874
are                 : 857
water               : 816
that                : 785
be                  : 777
which               : 722
;                   : 633
or                  : 581
from                : 551
but                 : 533
```

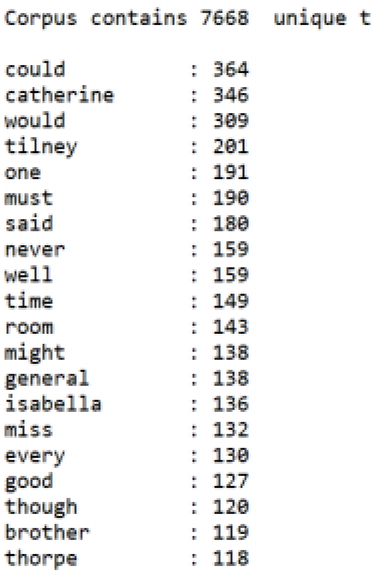
Document 4



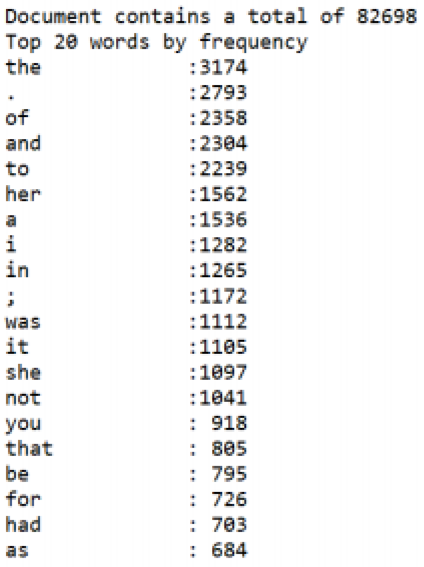
```
Corpus contains 4646  unique terms

Top 20 terms sorted by frequency
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
```



```
Corpus contains 4646  unique terms after Stemming.

Top 20 terms sorted by frequency
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
```

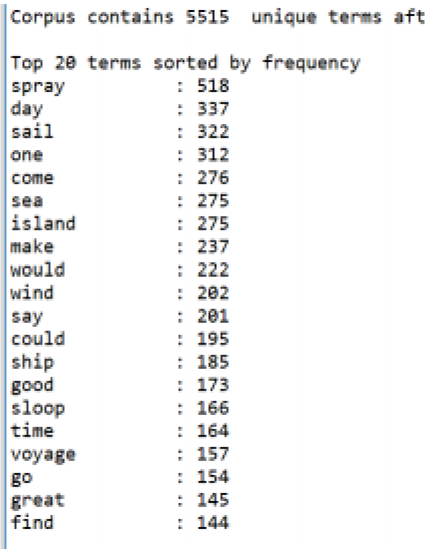


```
Corpus contains 7668  unique t

could           : 364
catherine       : 346
would           : 309
tilney          : 201
one             : 191
must            : 190
said            : 180
never           : 159
well            : 159
time            : 149
room            : 143
might           : 138
general         : 138
isabella        : 136
miss            : 132
every           : 130
good            : 127
though          : 120
brother         : 119
thorpe          : 118
```

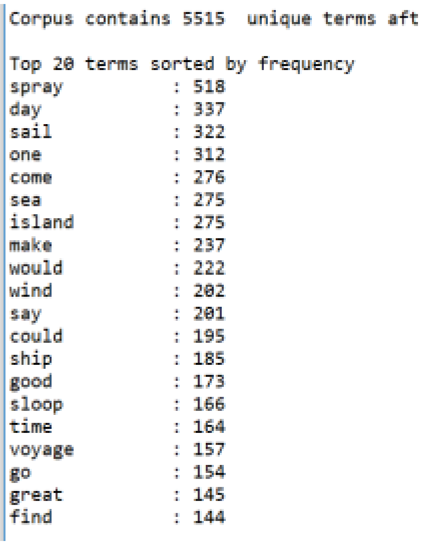


```
Document contains a total of 82698
Top 20 words by frequency
the             :3174
.               :2793
of              :2358
and             :2304
to              :2239
her             :1562
a               :1536
i               :1282
in              :1265
;               :1172
was             :1112
it              :1105
she             :1097
not             :1041
you             : 918
that            : 805
be              : 795
for             : 726
had             : 703
as              : 684
```

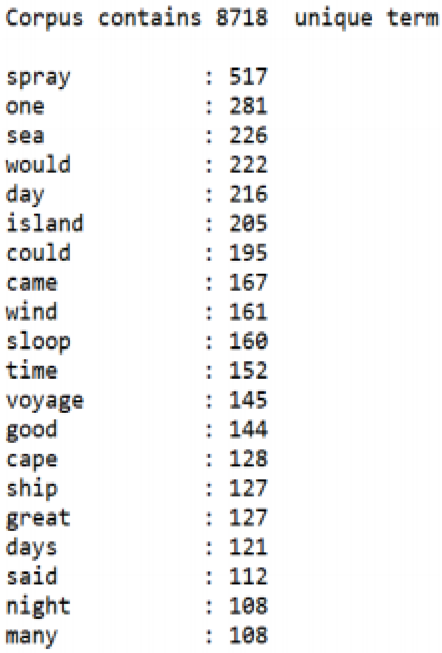
Document 5 – all 4 cases



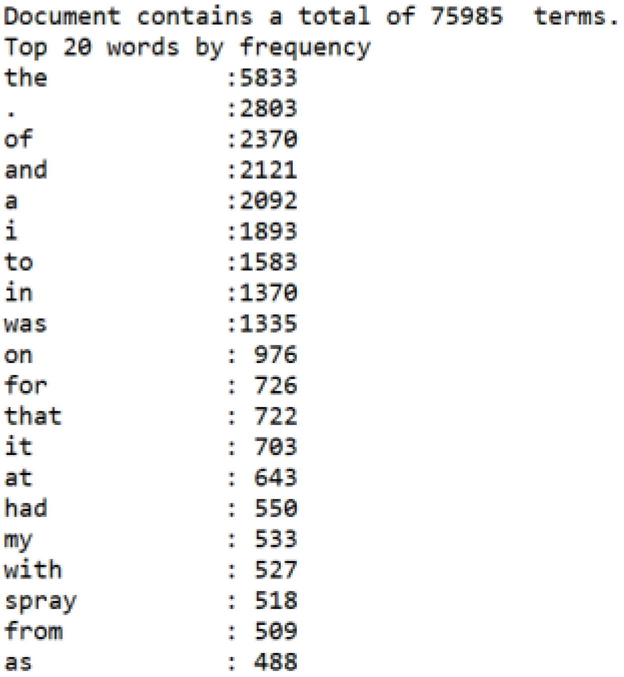
```
Corpus contains 5515  unique terms aft

Top 20 terms sorted by frequency
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
```



```
Corpus contains 5515  unique terms aft

Top 20 terms sorted by frequency
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
```

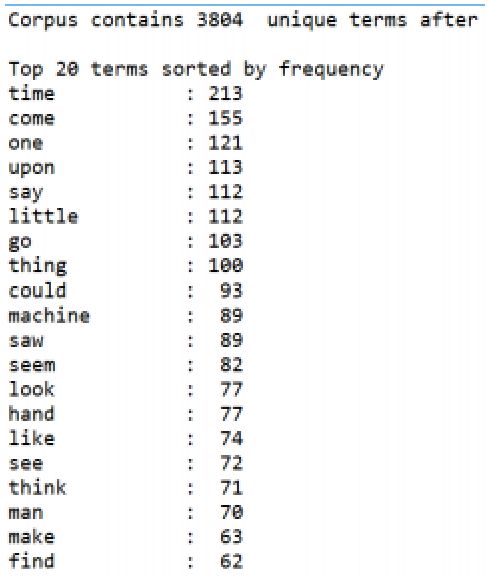


```
Corpus contains 8718  unique term

spray            : 517
one              : 281
sea              : 226
would            : 222
day              : 216
island           : 205
could            : 195
came             : 167
wind             : 161
sloop            : 160
time             : 152
voyage           : 145
good             : 144
cape             : 128
ship             : 127
great            : 127
days             : 121
said             : 112
night            : 108
many             : 108
```

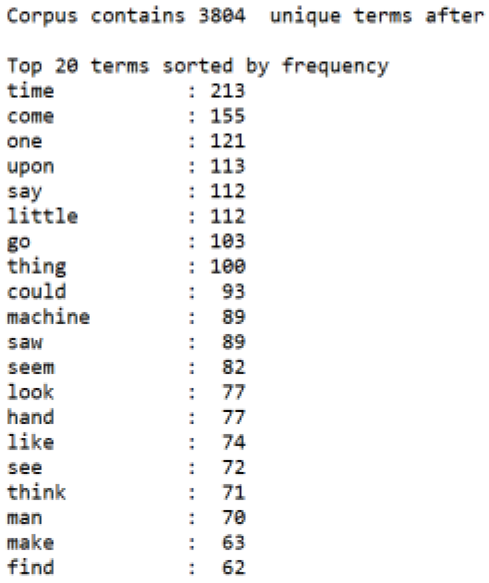


```
Document contains a total of 75985  terms.
Top 20 words by frequency
the                :5833
.                  :2803
of                 :2370
and                :2121
a                  :2092
i                  :1893
to                 :1583
in                 :1370
was                :1335
on                 : 976
for                : 726
that               : 722
it                 : 703
at                 : 643
had                : 550
my                 : 533
with               : 527
spray              : 518
from               : 509
as                 : 488
```

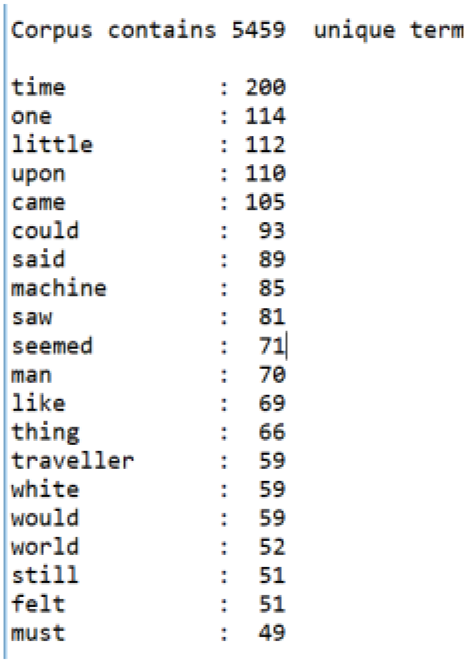
Document 6- all cases



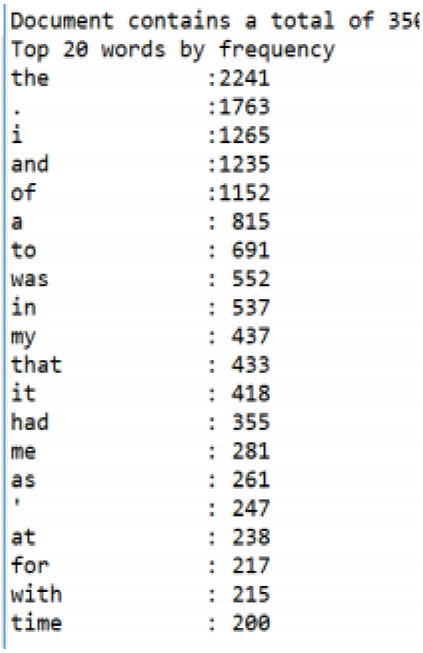
```
Corpus contains 3804  unique terms after

Top 20 terms sorted by frequency
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
```



```
Corpus contains 3804  unique terms after

Top 20 terms sorted by frequency
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
```

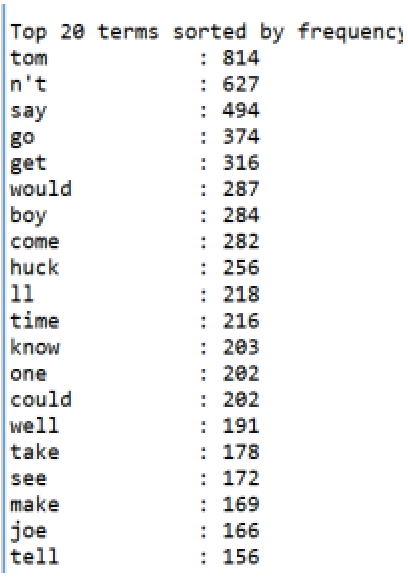


```
Corpus contains 5459  unique term

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           :  59
white               :  59
would               :  59
world               :  52
still               :  51
felt                :  51
must                :  49
```

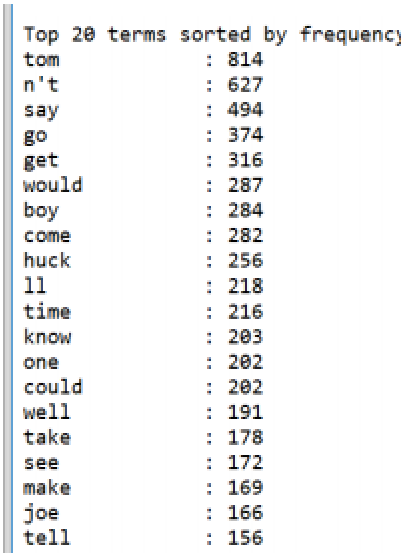


```
Document contains a total of 35(
Top 20 words by frequency
the                :2241
.                  :1763
i                  :1265
and                :1235
of                 :1152
a                  : 815
to                 : 691
was                : 552
in                 : 537
my                 : 437
that               : 433
it                 : 418
had                : 355
me                 : 281
as                 : 261
'                  : 247
at                 : 238
for                : 217
with               : 215
time               : 200
```

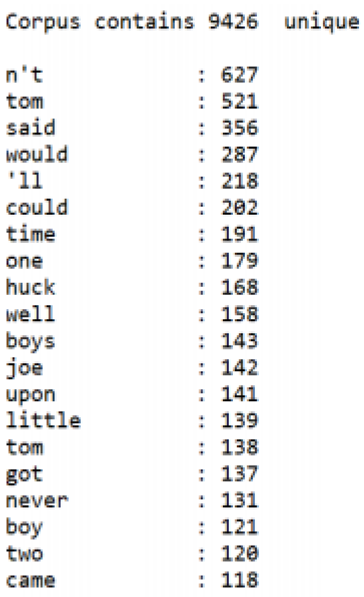
Document 7 – all 4 cases



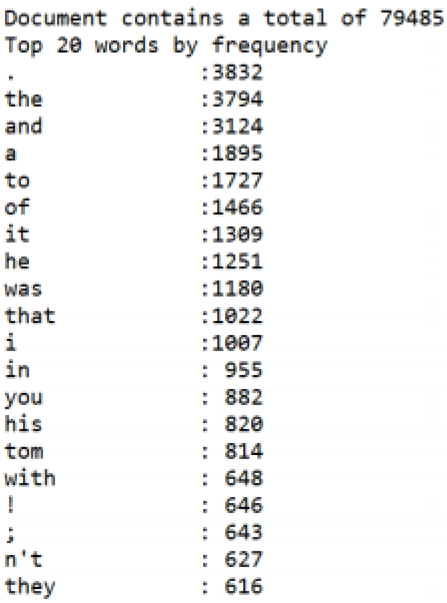
```
Top 20 terms sorted by frequency
tom                  : 814
n't                  : 627
say                  : 494
go                   : 374
get                  : 316
would                : 287
boy                  : 284
come                 : 282
huck                 : 256
ll                   : 218
time                 : 216
know                 : 203
one                  : 202
could                : 202
well                 : 191
take                 : 178
see                  : 172
make                 : 169
joe                  : 166
tell                 : 156
```



```
Top 20 terms sorted by frequency
tom                  : 814
n't                  : 627
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                 : 191
take                 : 178
see                  : 172
make                 : 169
joe                  : 166
tell                 : 156
```

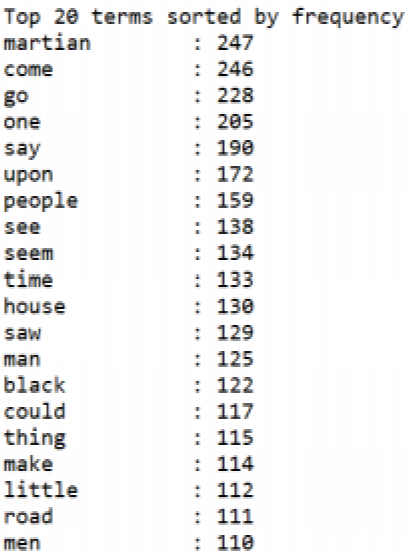


```
Corpus contains 9426  unique

n't                  : 627
tom                  : 521
said                 : 356
would                : 287
'll                  : 218
could                : 202
time                 : 191
one                  : 179
huck                 : 168
well                 : 158
boys                 : 143
joe                  : 142
upon                 : 141
little               : 139
tom                  : 138
got                  : 137
never                : 131
boy                  : 121
two                  : 120
came                 : 118
```

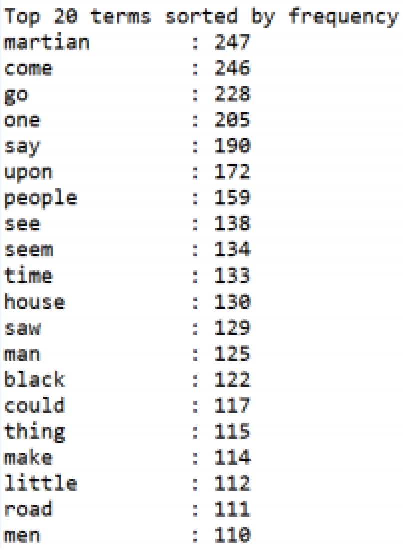


```
Document contains a total of 79485
Top 20 words by frequency
.                    :3832
the                  :3794
and                  :3124
a                    :1895
to                   :1727
of                   :1466
it                   :1309
he                   :1251
was                  :1180
that                 :1022
i                    :1007
in                   : 955
you                  : 882
his                  : 820
tom                  : 814
with                 : 648
!                    : 646
;                    : 643
n't                  : 627
they                 : 616
```

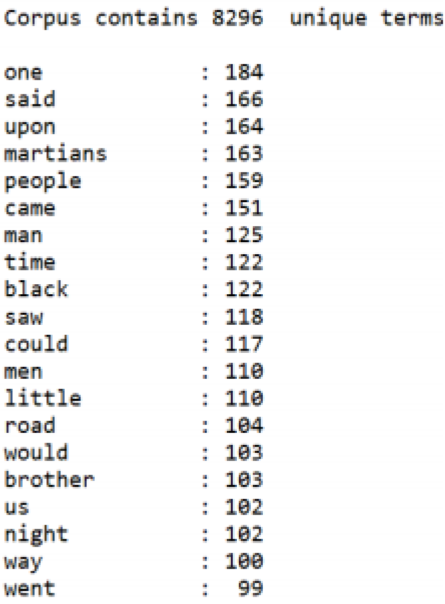
Document 8 – all 4 cases



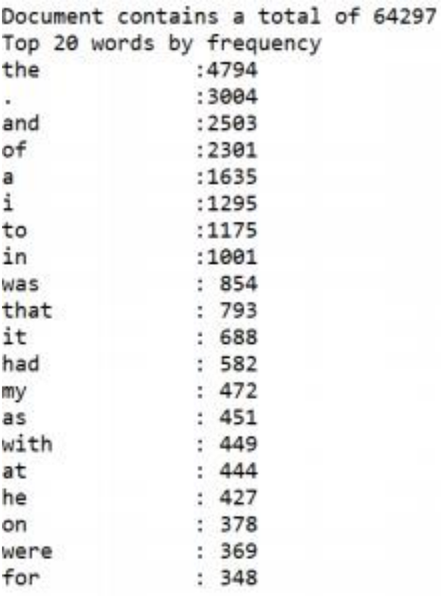
```
Top 20 terms sorted by frequency
martian         : 247
come            : 246
go              : 228
one             : 205
say             : 190
upon            : 172
people          : 159
see             : 138
seem            : 134
time            : 133
house           : 130
saw             : 129
man             : 125
black           : 122
could           : 117
thing           : 115
make            : 114
little          : 112
road            : 111
men             : 110
```



```
Top 20 terms sorted by frequency
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
```



```
Corpus contains 8296  unique terms

one             : 184
said            : 166
upon            : 164
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 a total of 64297
Top 20 words by frequency
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
```