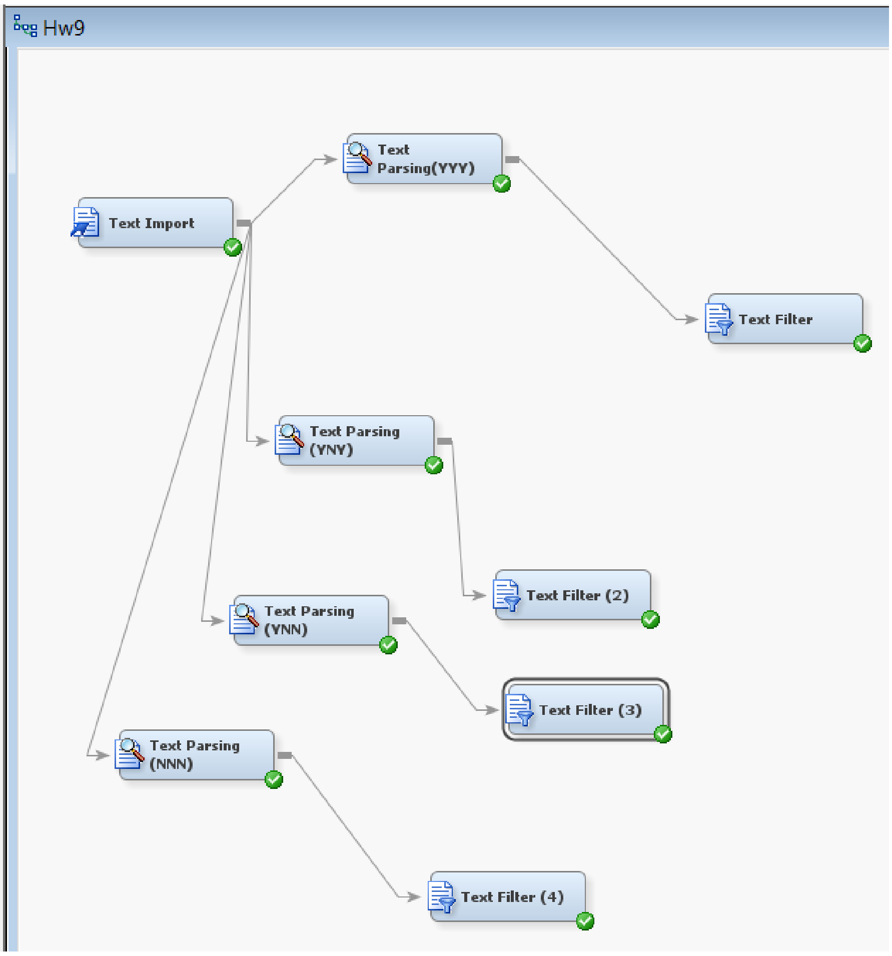
# STAT 656
# Week 9 Assignment

Name-Mayank Jaggi

UIN-526005299

# PART 1 SAS EM

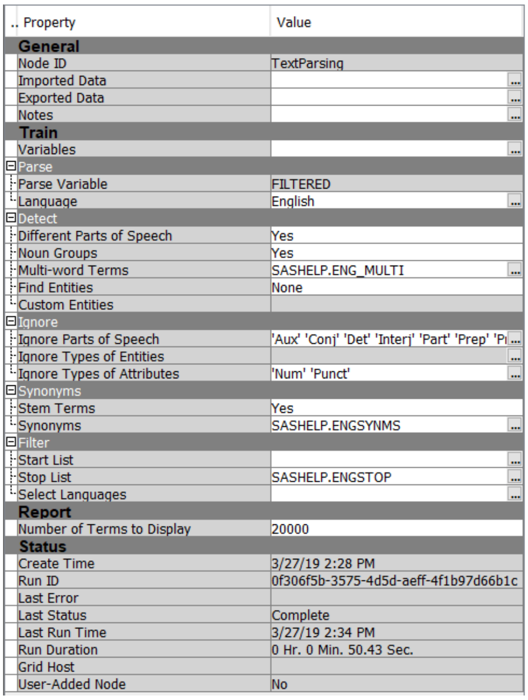
## 1) Project Diagram



## 2) Results

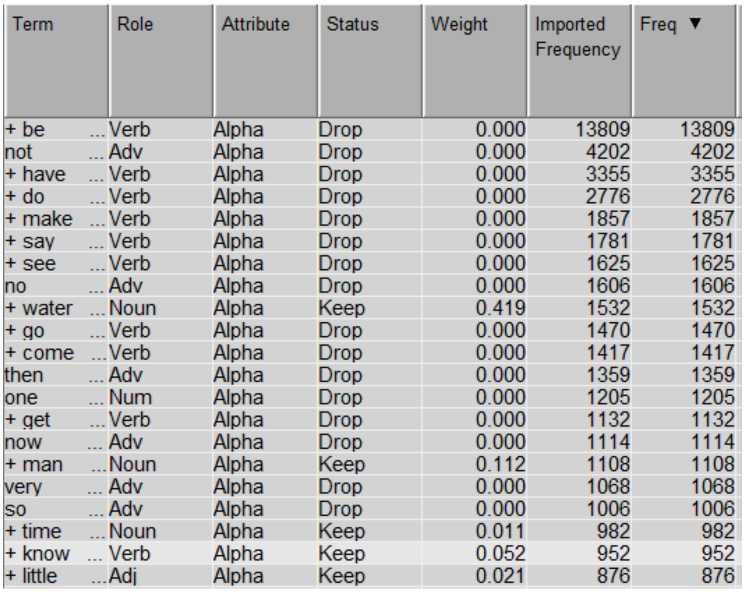
- **Remove Stop Words-Yes, POS-Yes, Stem-Yes**

Text Parsing Property



| .. 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' 'Prep' 'P |
| 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 | 3/27/19 2:28 PM |
| Run ID | 0f306f5b-3575-4d5d-aeff-4f1b97d66b1c |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 3/27/19 2:34 PM |
| Run Duration | 0 Hr. 0 Min. 50.43 Sec. |
| Grid Host | |
| User-Added Node | No |

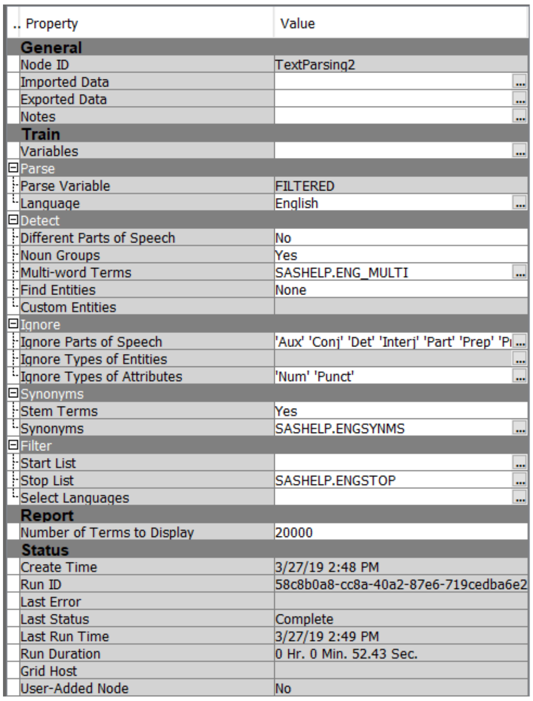
Result



| Term | Role | Attribute | Status | Weight | Imported Frequency | Freq ▼ |
|---|---|---|---|---|---|---|
| + be ... | Verb | Alpha | Drop | 0.000 | 13809 | 13809 |
| not ... | Adv | Alpha | Drop | 0.000 | 4202 | 4202 |
| + have ... | Verb | Alpha | Drop | 0.000 | 3355 | 3355 |
| + do ... | Verb | Alpha | Drop | 0.000 | 2776 | 2776 |
| + make ... | Verb | Alpha | Drop | 0.000 | 1857 | 1857 |
| + say ... | Verb | Alpha | Drop | 0.000 | 1781 | 1781 |
| + see ... | Verb | Alpha | Drop | 0.000 | 1625 | 1625 |
| no ... | Adv | Alpha | Drop | 0.000 | 1606 | 1606 |
| + water ... | Noun | Alpha | Keep | 0.419 | 1532 | 1532 |
| + go ... | Verb | Alpha | Drop | 0.000 | 1470 | 1470 |
| + come ... | Verb | Alpha | Drop | 0.000 | 1417 | 1417 |
| then ... | Adv | Alpha | Drop | 0.000 | 1359 | 1359 |
| one ... | Num | Alpha | Drop | 0.000 | 1205 | 1205 |
| + get ... | Verb | Alpha | Drop | 0.000 | 1132 | 1132 |
| now ... | Adv | Alpha | Drop | 0.000 | 1114 | 1114 |
| + man ... | Noun | Alpha | Keep | 0.112 | 1108 | 1108 |
| very ... | Adv | Alpha | Drop | 0.000 | 1068 | 1068 |
| so ... | Adv | Alpha | Drop | 0.000 | 1006 | 1006 |
| + time ... | Noun | Alpha | Keep | 0.011 | 982 | 982 |
| + know ... | Verb | Alpha | Keep | 0.052 | 952 | 952 |
| + little ... | Adj | Alpha | Keep | 0.021 | 876 | 876 |

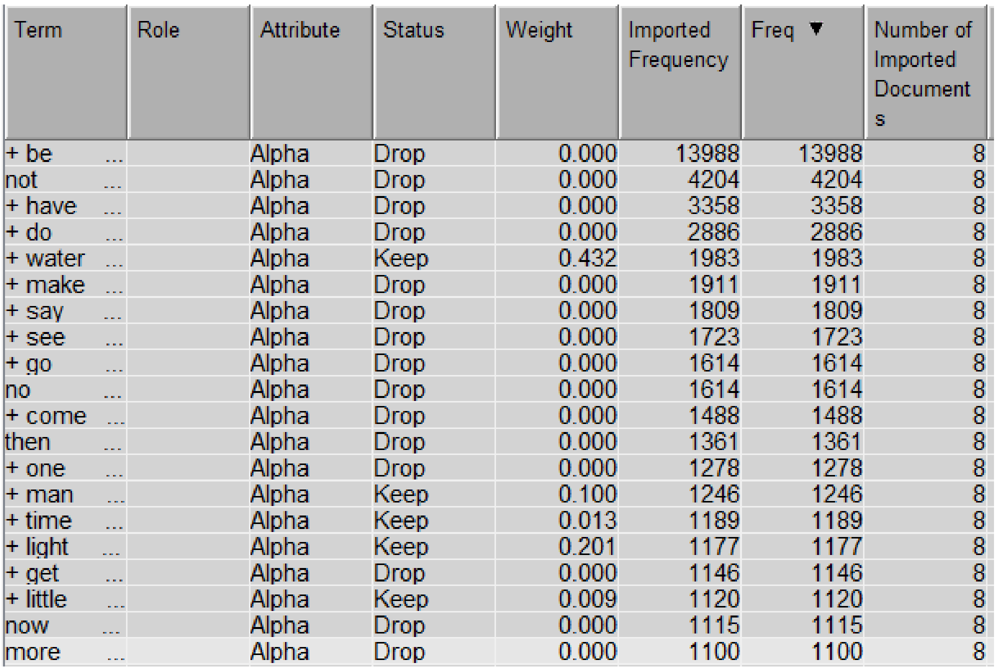
- **Remove Stop Words-Yes, POS-No,  Stem-Yes**

Text Parsing Property



| .. 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' 'Prep' 'Pı ... |
| 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 | 3/27/19 2:48 PM |
| Run ID | 58c8b0a8-cc8a-40a2-87e6-719cedba6e2 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 3/27/19 2:49 PM |
| Run Duration | 0 Hr. 0 Min. 52.43 Sec. |
| Grid Host | |
| User-Added Node | No |

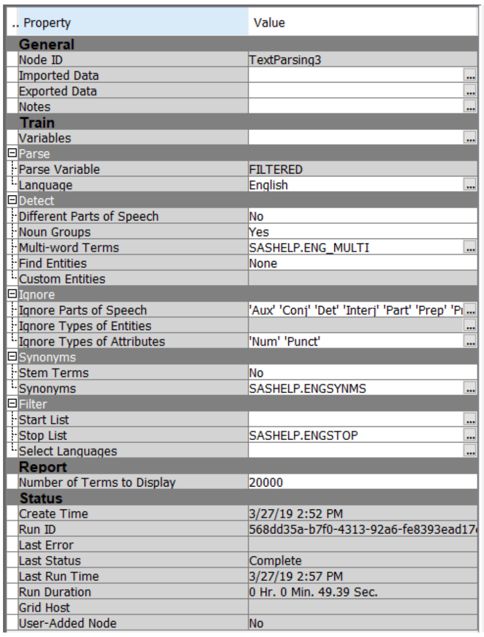
Result



| Term | Role | Attribute | Status | Weight | Imported Frequency | Freq ▼ | Number of Imported Documents |
|---|---|---|---|---|---|---|---|
| + be ... | | Alpha | Drop | 0.000 | 13988 | 13988 | 8 |
| not ... | | Alpha | Drop | 0.000 | 4204 | 4204 | 8 |
| + have ... | | Alpha | Drop | 0.000 | 3358 | 3358 | 8 |
| + do ... | | Alpha | Drop | 0.000 | 2886 | 2886 | 8 |
| + water ... | | Alpha | Keep | 0.432 | 1983 | 1983 | 8 |
| + make ... | | Alpha | Drop | 0.000 | 1911 | 1911 | 8 |
| + say ... | | Alpha | Drop | 0.000 | 1809 | 1809 | 8 |
| + see ... | | Alpha | Drop | 0.000 | 1723 | 1723 | 8 |
| + go ... | | Alpha | Drop | 0.000 | 1614 | 1614 | 8 |
| no ... | | Alpha | Drop | 0.000 | 1614 | 1614 | 8 |
| + come ... | | Alpha | Drop | 0.000 | 1488 | 1488 | 8 |
| then ... | | Alpha | Drop | 0.000 | 1361 | 1361 | 8 |
| + one ... | | Alpha | Drop | 0.000 | 1278 | 1278 | 8 |
| + man ... | | Alpha | Keep | 0.100 | 1246 | 1246 | 8 |
| + time ... | | Alpha | Keep | 0.013 | 1189 | 1189 | 8 |
| + light ... | | Alpha | Keep | 0.201 | 1177 | 1177 | 8 |
| + get ... | | Alpha | Drop | 0.000 | 1146 | 1146 | 8 |
| + little ... | | Alpha | Keep | 0.009 | 1120 | 1120 | 8 |
| now ... | | Alpha | Drop | 0.000 | 1115 | 1115 | 8 |
| more ... | | Alpha | Drop | 0.000 | 1100 | 1100 | 8 |

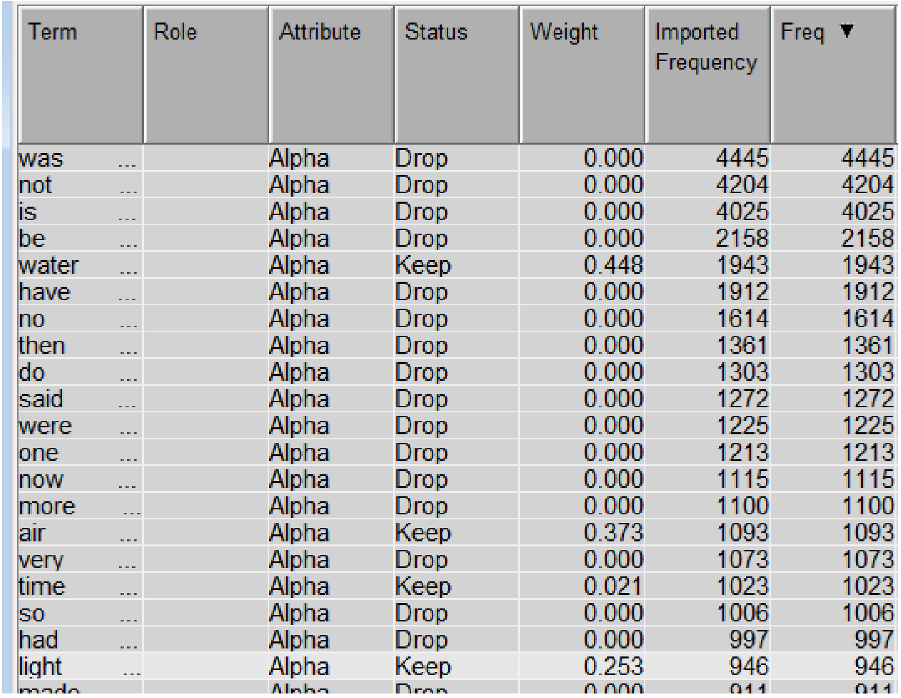
- **Remove Stop Words-Yes, POS-No, Stem-No**

Text Parsing Property



| .. 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' 'Prep' '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 | 3/27/19 2:52 PM |
| Run ID | 568dd35a-b7f0-4313-92a6-fe8393ead17 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 3/27/19 2:57 PM |
| Run Duration | 0 Hr. 0 Min. 49.39 Sec. |
| Grid Host | |
| User-Added Node | No |

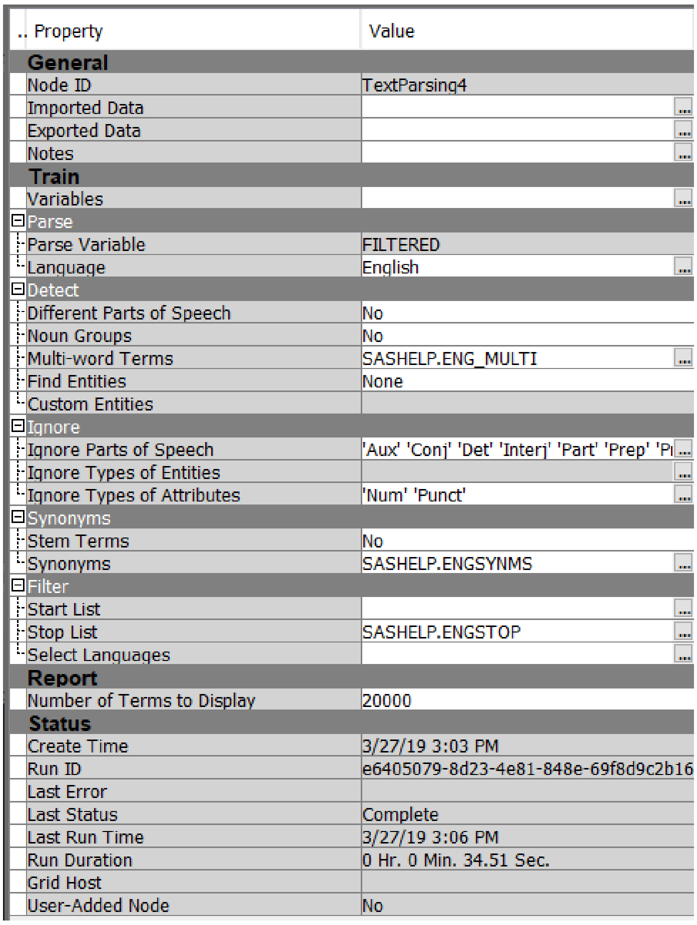
Result



| Term | Role | Attribute | Status | Weight | Imported Frequency | Freq ▼ |
|---|---|---|---|---|---|---|
| was | | Alpha | Drop | 0.000 | 4445 | 4445 |
| not | | Alpha | Drop | 0.000 | 4204 | 4204 |
| is | | Alpha | Drop | 0.000 | 4025 | 4025 |
| be | | Alpha | Drop | 0.000 | 2158 | 2158 |
| water | | Alpha | Keep | 0.448 | 1943 | 1943 |
| have | | Alpha | Drop | 0.000 | 1912 | 1912 |
| no | | Alpha | Drop | 0.000 | 1614 | 1614 |
| then | | Alpha | Drop | 0.000 | 1361 | 1361 |
| do | | Alpha | Drop | 0.000 | 1303 | 1303 |
| said | | Alpha | Drop | 0.000 | 1272 | 1272 |
| were | | Alpha | Drop | 0.000 | 1225 | 1225 |
| one | | Alpha | Drop | 0.000 | 1213 | 1213 |
| now | | Alpha | Drop | 0.000 | 1115 | 1115 |
| more | | Alpha | Drop | 0.000 | 1100 | 1100 |
| air | | Alpha | Keep | 0.373 | 1093 | 1093 |
| very | | Alpha | Drop | 0.000 | 1073 | 1073 |
| time | | Alpha | Keep | 0.021 | 1023 | 1023 |
| so | | Alpha | Drop | 0.000 | 1006 | 1006 |
| had | | Alpha | Drop | 0.000 | 997 | 997 |
| light | | Alpha | Keep | 0.253 | 946 | 946 |
| made | | Alpha | Drop | 0.000 | 911 | 911 |

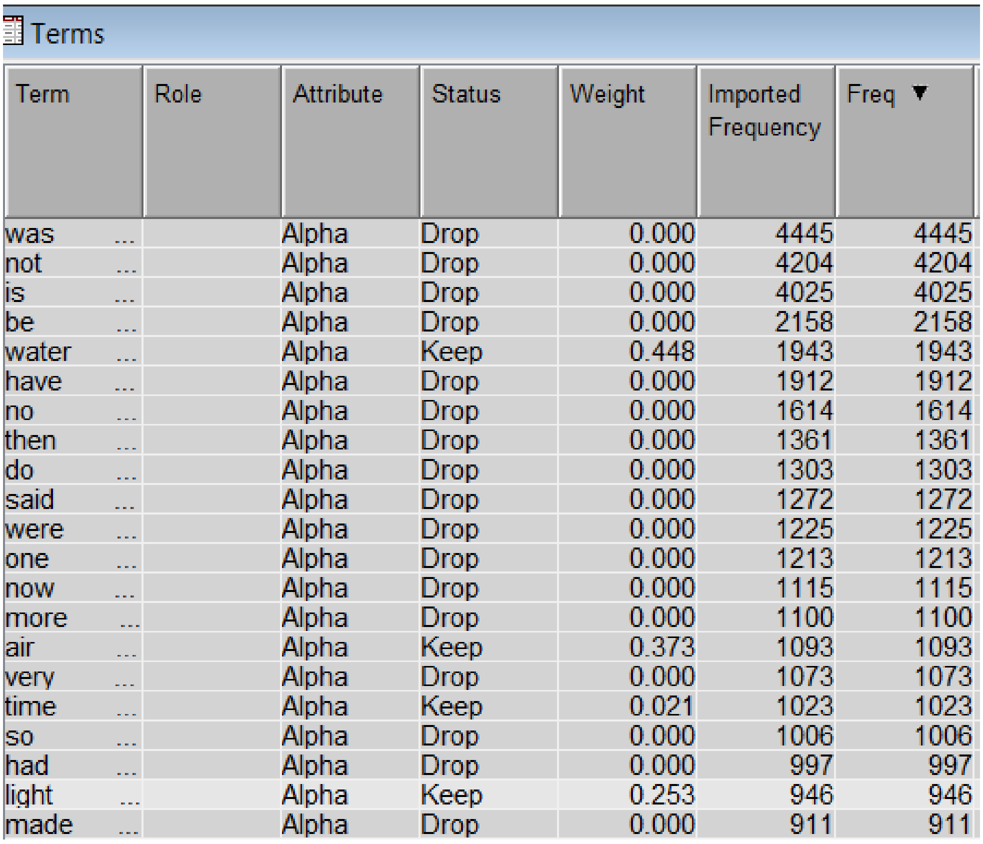
- **Remove Stop Words-No, POS-No, Stem-No**

Text Parsing Property



| .. Property | Value |
|---|---|
| **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 | No |
| Multi-word Terms | SASHELP.ENG_MULTI |
| Find Entities | None |
| Custom Entities | |
| ⊟Ignore | |
| Ignore Parts of Speech | 'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Prep' 'P |
| 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 | 3/27/19 3:03 PM |
| Run ID | e6405079-8d23-4e81-848e-69f8d9c2b16 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 3/27/19 3:06 PM |
| Run Duration | 0 Hr. 0 Min. 34.51 Sec. |
| Grid Host | |
| User-Added Node | No |

Result



Terms

| Term | Role | Attribute | Status | Weight | Imported Frequency | Freq ▼ |
|---|---|---|---|---|---|---|
| was | | Alpha | Drop | 0.000 | 4445 | 4445 |
| not | | Alpha | Drop | 0.000 | 4204 | 4204 |
| is | | Alpha | Drop | 0.000 | 4025 | 4025 |
| be | | Alpha | Drop | 0.000 | 2158 | 2158 |
| water | | Alpha | Keep | 0.448 | 1943 | 1943 |
| have | | Alpha | Drop | 0.000 | 1912 | 1912 |
| no | | Alpha | Drop | 0.000 | 1614 | 1614 |
| then | | Alpha | Drop | 0.000 | 1361 | 1361 |
| do | | Alpha | Drop | 0.000 | 1303 | 1303 |
| said | | Alpha | Drop | 0.000 | 1272 | 1272 |
| were | | Alpha | Drop | 0.000 | 1225 | 1225 |
| one | | Alpha | Drop | 0.000 | 1213 | 1213 |
| now | | Alpha | Drop | 0.000 | 1115 | 1115 |
| more | | Alpha | Drop | 0.000 | 1100 | 1100 |
| air | | Alpha | Keep | 0.373 | 1093 | 1093 |
| very | | Alpha | Drop | 0.000 | 1073 | 1073 |
| time | | Alpha | Keep | 0.021 | 1023 | 1023 |
| so | | Alpha | Drop | 0.000 | 1006 | 1006 |
| had | | Alpha | Drop | 0.000 | 997 | 997 |
| light | | Alpha | Keep | 0.253 | 946 | 946 |
| made | | Alpha | Drop | 0.000 | 911 | 911 |

```
## Part 2 Python

##a) Python Program

# -*- coding: utf-8 -*-
"""
Created on Wed Mar 27 10:08:22 2019

@author: mayank
"""

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 string
import pandas as pd

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('stopwords')
nltk.download('wordnet')

def textanalytics(scenario):

   # x=[]
    #Intialize a dictionary for words extracted
    counter={}

    # Read Document
    file_path="D:\\Work\\Course Work\\Semester 4\\STAT 656\\Lectures & 
Assignment\\Week 9\\Week 9 Assignment\\TextFiles"
    files=['T1.txt','T2.txt','T3.txt','T4.txt','T5.txt','T6.txt','T7.txt','T8.txt']


    for text in files:
        with open (file_path+"\\"+text, "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('-', ' ')
        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) and \
        (word!='description') and (word !='dtype') \
        and (word != 'object') and (word!="'s")]

        #print(text,"Document contains a total of", len(tokens), " terms.\n")
        #print(len(tokens))

        # POS Tagging
        if scenario==1:
            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))
            #print(tagged_tokens)

        # Remove stop words
        if scenario==1:
            stop = stopwords.words('english') + list(string.punctuation)
            stop_tokens = [word for word in tagged_tokens if word[0] not in stop]
        elif scenario==2 or scenario==3:
            stop = stopwords.words('english') + list(string.punctuation)
            stop_tokens = [word for word in tokens if word not in stop]

        if scenario==1 or scenario==2 or scenario==3:

            # 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(text,"\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))
            #print(stop_tokens)
#        if scenario==3:
#            #x.append(stop_tokens)

        #Stemming
        if scenario==1 or scenario==2:
            # 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))
                # Get token distribution
            fdist = FreqDist(stemmed_tokens)
            #x.append(stemmed_tokens)
            #print(x)
            #print(text,"\nCorpus contains", len(fdist.items()), \
            #" unique terms after Stemming.\n")
#            print(stemmed_tokens)
#            print(fdist)
            #print(x)




        if scenario==1 or scenario==2:
            counter[text]=len(fdist.items())
        elif scenario==3:
            counter[text]=len(token_dist.items())
        else:
            counter[text]=len(tokens)

    print("Scenario No: ",scenario,"\n", "Words extracted from each file:
\n",counter,\
          "\n\nTotal Number of terms extracted from all files:
",sum(counter.values()),"\n")




#    if scenario==1 or scenario==2:
#        a=FreqDist(x)
#        print(a)
#
#        for word, freq in a.most_common(20):
```

```
#            a[word]=freq
#            print('{:<15s}:{:>4d}'.format(word, freq))
##            fdist_top = nltk.probability.FreqDist()
#
#    elif scenario==3:
#        b=FreqDist(x)
#        for word, freq in b.most_common(20):
#            print('{:<15s}:{:>4d}'.format(word, freq))
##            fdist_top = nltk.probability.FreqDist()



textanalytics(1)
textanalytics(2)
textanalytics(3)
textanalytics(4)


###################PARTIAL OUTPUT########################
```

```
##b)

Scenario No:  1
 Words extracted from each file:
 {'T1.txt': 6375, 'T2.txt': 3981, 'T3.txt': 5282, 'T4.txt': 4646, 'T5.txt': 5515,
'T6.txt': 3804, 'T7.txt': 5789, 'T8.txt': 5407}

Total Number of terms extracted from all files:  40799

Scenario No:  2
 Words extracted from each file:
 {'T1.txt': 29, 'T2.txt': 29, 'T3.txt': 29, 'T4.txt': 25, 'T5.txt': 29, 'T6.txt':
28, 'T7.txt': 28, 'T8.txt': 26}

Total Number of terms extracted from all files:  223

Scenario No:  3
 Words extracted from each file:
 {'T1.txt': 8012, 'T2.txt': 5221, 'T3.txt': 6724, 'T4.txt': 5911, 'T5.txt': 6792,
'T6.txt': 4552, 'T7.txt': 7231, 'T8.txt': 6610}

Total Number of terms extracted from all files:  51053

Scenario No:  4
 Words extracted from each file:
 {'T1.txt': 86120, 'T2.txt': 108341, 'T3.txt': 104654, 'T4.txt': 82698, 'T5.txt':
75985, 'T6.txt': 35066, 'T7.txt': 79485, 'T8.txt': 64297}

Total Number of terms extracted from all files:  636646
```

In Scenario 2 stemming doesnt make sense without POS as POS tags each word with its corressponding part of speech and stemming uses the pos of a word to convert it to present tense. The counter should be calculated after 'removing stop words' step.

The corresponding results would be:

```
Scenario No:  2
 Words extracted from each file:
 {'T1.txt': 8012, 'T2.txt': 5221, 'T3.txt': 6724, 'T4.txt': 5911, 'T5.txt': 6792,
'T6.txt': 4552, 'T7.txt': 7231, 'T8.txt': 6610}

Total Number of terms extracted from all files:  51053
```