STAT 656

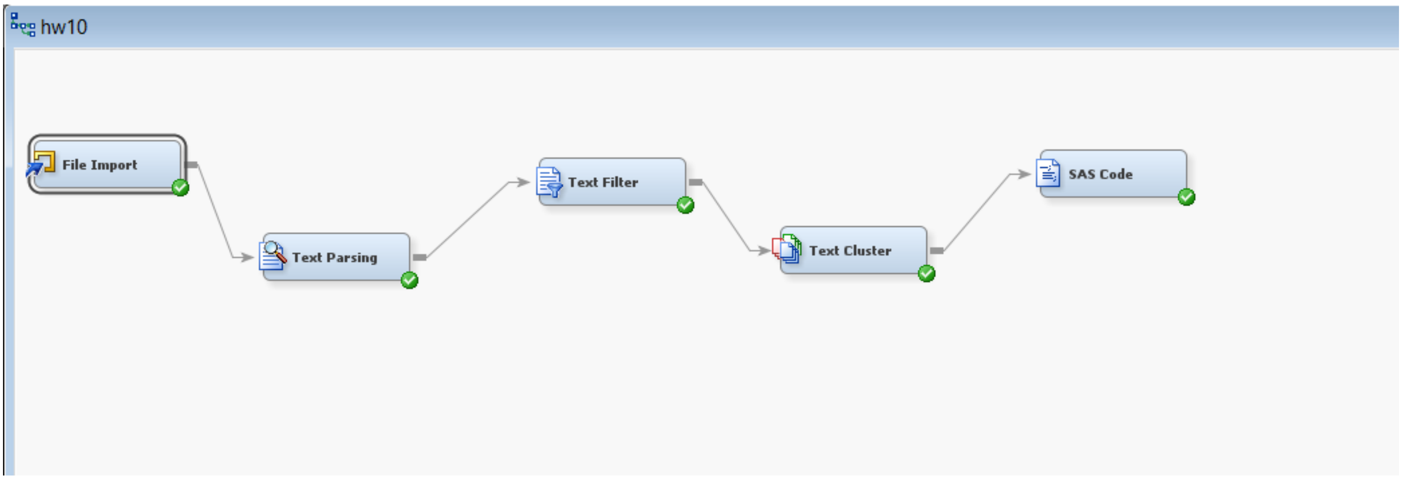
Week 10 Assignment

Name-Mayank Jaggi

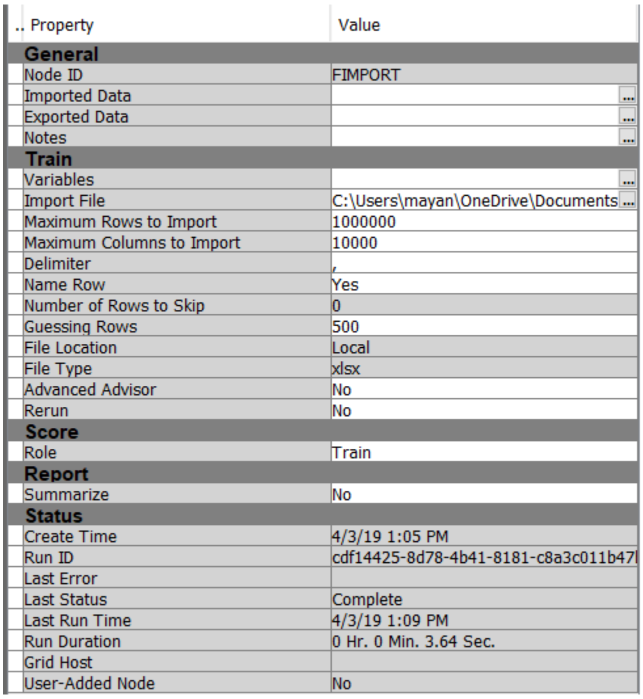
UIN-526005299

PART 1 SAS EM

1) Project Diagram & Property Window

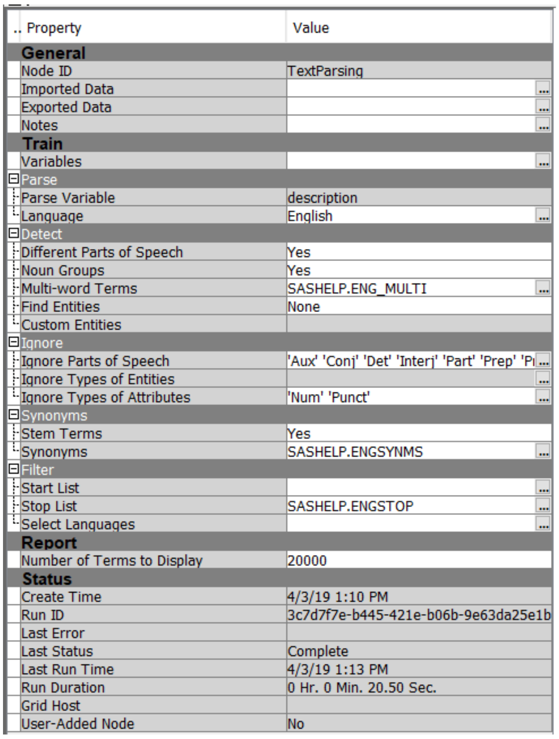


File Import Property



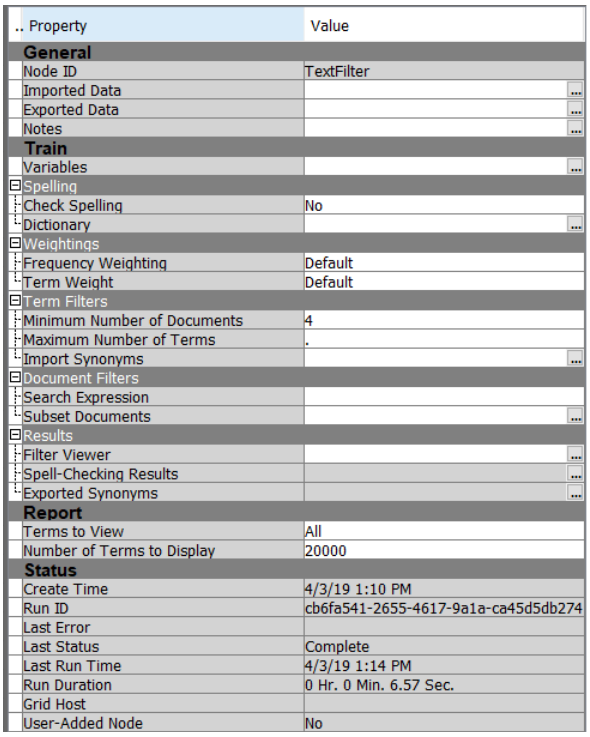
.. Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	C:\Users\mayan\OneDrive\Documents ...
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advisor	No
Rerun	No
Score	
Role	Train
Report	
Summarize	No
Status	
Create Time	4/3/19 1:05 PM
Run ID	cdf14425-8d78-4b41-8181-c8a3c011b47
Last Error	
Last Status	Complete
Last Run Time	4/3/19 1:09 PM
Run Duration	0 Hr. 0 Min. 3.64 Sec.
Grid Host	
User-Added Node	No

Text Parsing Property



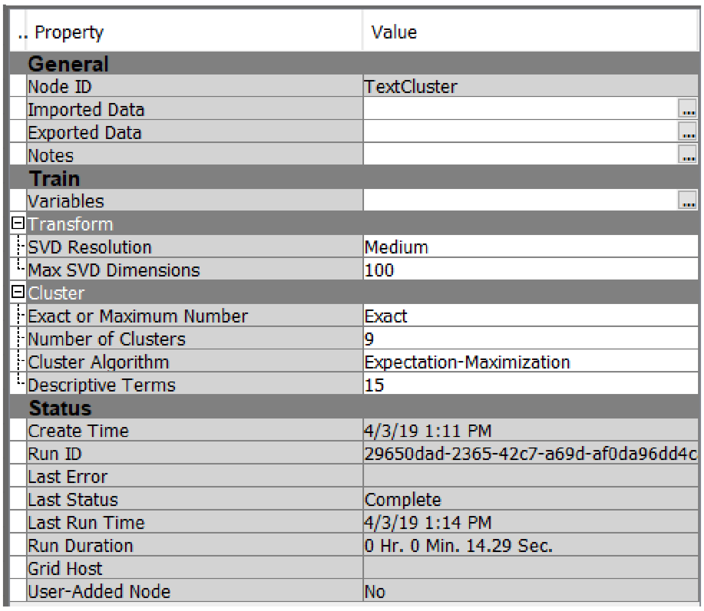
.. Property	Value
General	
Node ID	TextParsing
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Parse	
Parse Variable	description
Language	English ...
⊟ Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI ...
Find Entities	None
Custom Entities	
⊟ Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part' 'Prep' '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	4/3/19 1:10 PM
Run ID	3c7d7f7e-b445-421e-b06b-9e63da25e1b
Last Error	
Last Status	Complete
Last Run Time	4/3/19 1:13 PM
Run Duration	0 Hr. 0 Min. 20.50 Sec.
Grid Host	
User-Added Node	No

Text Filter Property



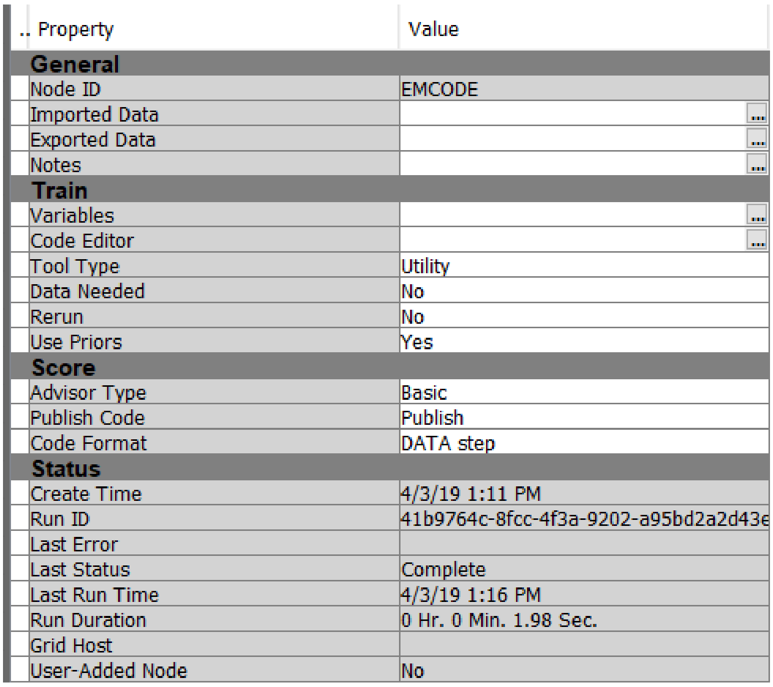
.. Property	Value
General	
Node ID	TextFilter
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Spelling	
Check Spelling	No
Dictionary	...
⊟ Weightings	
Frequency Weighting	Default
Term Weight	Default
⊟ Term Filters	
Minimum Number of Documents	4
Maximum Number of Terms	.
Import Synonyms	...
⊟ Document Filters	
Search Expression	
Subset Documents	...
⊟ Results	
Filter Viewer	...
Spell-Checking Results	...
Exported Synonyms	...
Report	
Terms to View	All
Number of Terms to Display	20000
Status	
Create Time	4/3/19 1:10 PM
Run ID	cb6fa541-2655-4617-9a1a-ca45d5db274
Last Error	
Last Status	Complete
Last Run Time	4/3/19 1:14 PM
Run Duration	0 Hr. 0 Min. 6.57 Sec.
Grid Host	
User-Added Node	No

Text Cluster Property



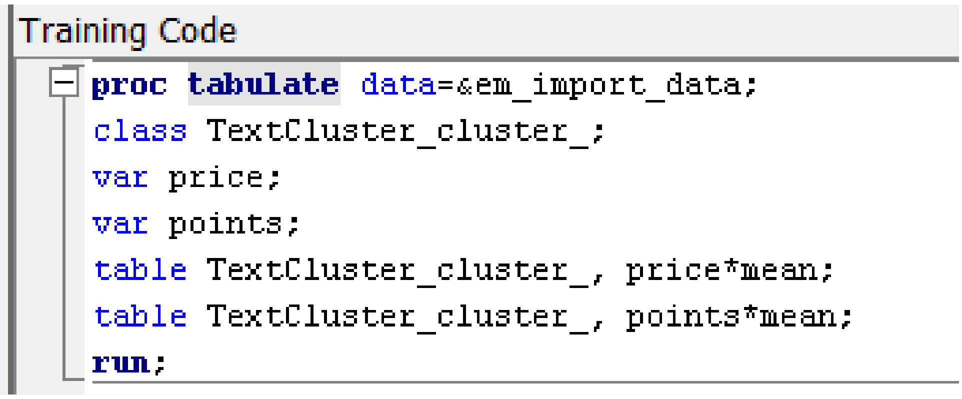
.. Property	Value
General	
Node ID	TextCluster
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Transform	
SVD Resolution	Medium
Max SVD Dimensions	100
⊟ Cluster	
Exact or Maximum Number	Exact
Number of Clusters	9
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	4/3/19 1:11 PM
Run ID	29650dad-2365-42c7-a69d-af0da96dd4c
Last Error	
Last Status	Complete
Last Run Time	4/3/19 1:14 PM
Run Duration	0 Hr. 0 Min. 14.29 Sec.
Grid Host	
User-Added Node	No

SAS Code Property



.. Property	Value
General	
Node ID	EMCODE
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Code Editor	...
Tool Type	Utility
Data Needed	No
Rerun	No
Use Priors	Yes
Score	
Advisor Type	Basic
Publish Code	Publish
Code Format	DATA step
Status	
Create Time	4/3/19 1:11 PM
Run ID	41b9764c-8fcc-4f3a-9202-a95bd2a2d43e
Last Error	
Last Status	Complete
Last Run Time	4/3/19 1:16 PM
Run Duration	0 Hr. 0 Min. 1.98 Sec.
Grid Host	
User-Added Node	No

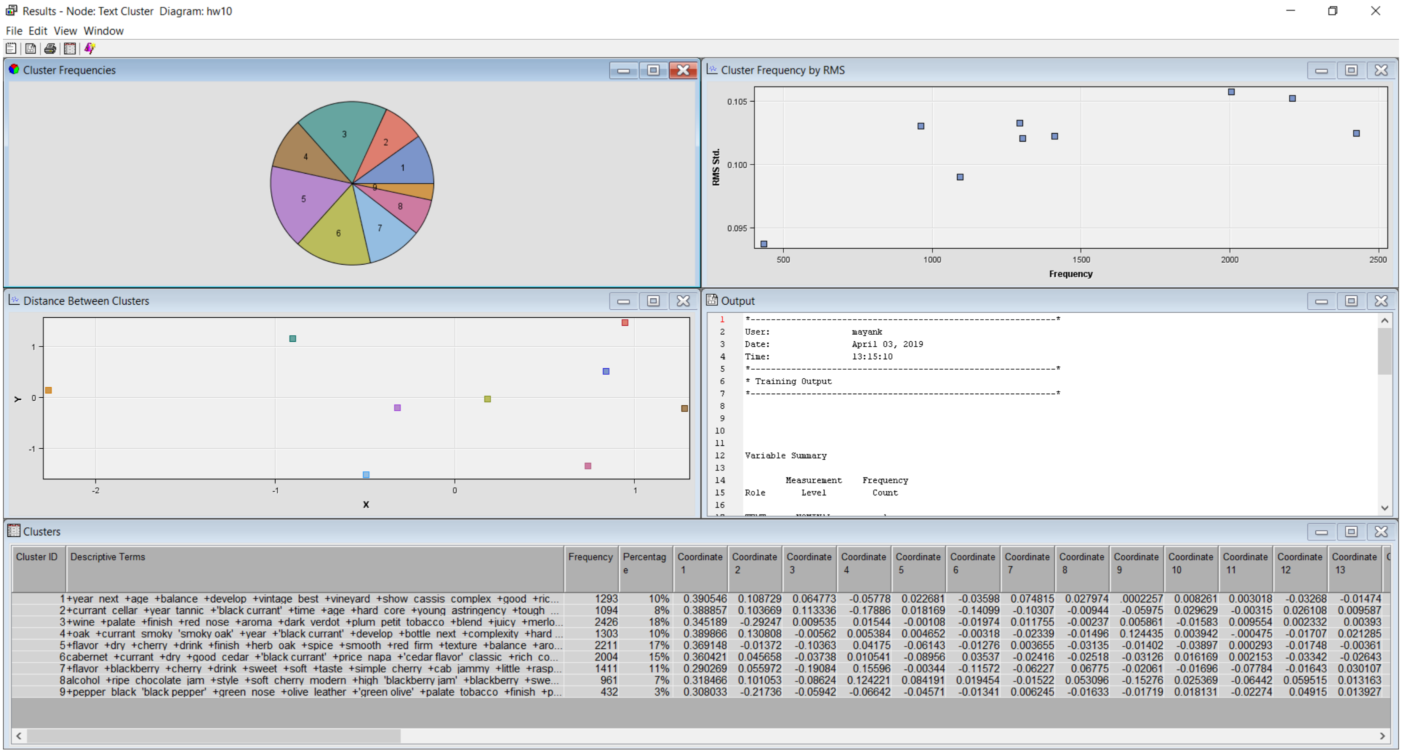
SAS Code



Training Code

```
proc tabulate data=&em_import_data;
class TextCluster_cluster_;
var price;
var points;
table TextCluster_cluster_, price*mean;
table TextCluster_cluster_, points*mean;
run;
```

2) Results



```
*------------------------------------------------------------*
User:                mayank
Date:                April 03, 2019
Time:                13:15:10
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
Role        Level          Count

```

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coordinate 7	Coordinate 8	Coordinate 9	Coordinate 10	Coordinate 11	Coordinate 12	Coordinate 13
1	+year next +age +balance +develop +vintage best +vineyard +show cassis complex +good +ric...	1293	10%	0.390546	0.108729	0.064773	-0.05778	0.022681	-0.03598	0.074815	0.027974	.0002257	0.008261	0.003018	-0.03268	-0.01474
2	+currant cellar +year tannic +'black currant' +time +age +hard core +young astringency +tough ...	1094	8%	0.388857	0.103669	0.113336	-0.17886	0.018169	-0.14099	-0.10307	-0.00944	-0.05975	0.029629	-0.00315	0.026108	0.009587
3	+wine +palate +finish +red nose +aroma +dark verdot +plum petit tobacco +blend +juicy +merlo...	2426	18%	0.345189	-0.29247	0.009535	0.01544	-0.00108	-0.01974	0.011755	-0.00237	0.005861	-0.01583	0.009554	0.002332	0.00393
4	+oak +currant smoky 'smoky oak' +year +'black currant' +develop +bottle next +complexity +hard ...	1303	10%	0.389866	0.130808	-0.00562	0.005384	0.004652	-0.00318	-0.02339	-0.01496	0.124435	0.003942	-.000475	-0.01707	0.021285
5	+flavor +dry +cherry +drink +finish +herb oak +spice +smooth +red firm +texture +balance +aro...	2211	17%	0.369148	-0.01372	-0.10363	0.04175	-0.06143	-0.01276	0.003655	-0.03135	-0.01402	-0.03897	0.000293	-0.01748	-0.00361
6	cabernet +currant +dry +good cedar +'black currant' +price napa +'cedar flavor' classic +rich co...	2004	15%	0.360421	0.045658	-0.03738	0.010541	-0.08956	0.03537	-0.02416	-0.02518	-0.03126	0.016169	0.002153	-0.03342	-0.02643
7	+flavor +blackberry +cherry +drink +sweet +soft +taste +simple cherry +cab jammy +little +rasp...	1411	11%	0.290269	0.055972	-0.19084	0.15596	-0.00344	-0.11572	-0.06227	0.06775	-0.02061	-0.01696	-0.07784	-0.01643	0.030107
8	alcohol +ripe chocolate jam +style +soft cherry modern +high 'blackberry jam' +blackberry +swe...	961	7%	0.318466	0.101053	-0.08624	0.124221	0.084191	0.019454	-0.01522	0.053096	-0.15276	0.025369	-0.06442	0.059515	0.013163
9	+pepper black 'black pepper' +green nose +olive leather +'green olive' +palate tobacco +finish +p...	432	3%	0.308033	-0.21736	-0.05942	-0.06642	-0.04571	-0.01341	0.006245	-0.01633	-0.01719	0.018131	-0.02274	0.04915	0.013927

```
*-------------------------------------------------------------*
User:                mayank
Date:                April 03, 2019
Time:                13:16:59
*-------------------------------------------------------------*
* Training Output
*-------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

ID           NOMINAL            1
INPUT        INTERVAL          79
INPUT        NOMINAL            2
REJECTED     INTERVAL           9
SEGMENT      NOMINAL            1
TEXT         NOMINAL            1




```

```
-----------------------------------------------------------------------------
|                                                              |   price    |
|                                                              |------------|
|                                                              |    Mean    |
|--------------------------------------------------------------+------------|
|TextCluster_cluster_                                          |            |
|--------------------------------------------------------------|            |
|1                                                             |       73.83|
|--------------------------------------------------------------+------------|
|2                                                             |       73.87|
|--------------------------------------------------------------+------------|
|3                                                             |       63.27|
|--------------------------------------------------------------+------------|
|4                                                             |       57.53|
|--------------------------------------------------------------+------------|
|5                                                             |       45.98|
|--------------------------------------------------------------+------------|
|6                                                             |       45.56|
|--------------------------------------------------------------+------------|
|7                                                             |       32.24|
|--------------------------------------------------------------+------------|
|8                                                             |       60.29|
|--------------------------------------------------------------+------------|
|9                                                             |       59.79|
-----------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
|                                                               |   points   |
|                                                               |------------|
|                                                               |    Mean    |
|---------------------------------------------------------------+------------|
|TextCluster_cluster_                                           |            |
|---------------------------------------------------------------|            |
|1                                                              |       91.21|
|---------------------------------------------------------------+------------|
|2                                                              |       90.68|
|---------------------------------------------------------------+------------|
|3                                                              |       89.58|
|---------------------------------------------------------------+------------|
|4                                                              |       89.43|
|---------------------------------------------------------------+------------|
|5                                                              |       88.42|
|---------------------------------------------------------------+------------|
|6                                                              |       88.47|
|---------------------------------------------------------------+------------|
|7                                                              |       84.38|
|---------------------------------------------------------------+------------|
|8                                                              |       87.70|
|---------------------------------------------------------------+------------|
|9                                                              |       88.88|
------------------------------------------------------------------------------


*-------------------------------------------------------------*
* Score Output
*-------------------------------------------------------------*


*-------------------------------------------------------------*
* Report Output
*-------------------------------------------------------------*
```

PART 2 PYTHON

1) PYTHON CODE

```
# -*- coding: utf-8 -*-
"""
Created on Wed Apr  3 12:29:38 2019

@author: mayank
"""

import pandas as pd
import numpy as np
import string
import nltk
from nltk import pos_tag
from nltk.tokenize import word_tokenize
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet as wn
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.decomposition import LatentDirichletAllocation


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


def analyzer_func(s):
    # List of synonyms
    syns = {'veh': 'vehicle', 'car': 'vehicle', 'chev':'cheverolet', \
              'chevy':'cheverolet', 'air bag': 'airbag', \
              'seat belt':'seatbelt', "n't":'not', 'to30':'to 30', \
              'wont':'would not', 'cant':'can not', 'cannot':'can not', \
              'couldnt':'could not', 'shouldnt':'should not', \
              'wouldnt':'would not', }


    s = s.lower()
    s = s.replace(',', '. ')

    tokens = word_tokenize(s)
    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")]


    for i in range(len(tokens)):
        if tokens[i] in syns:
            tokens[i] = syns[tokens[i]]

    # Removing stop words
    punctuation = list(string.punctuation)+['..', '...']
    pronouns = ['i', 'he', 'she', 'it', 'him', 'they', 'we', 'us', 'them']
    stop = stopwords.words('english') + punctuation + pronouns
    filtered_terms = [word for word in tokens if (word not in stop) and \
                  (len(word)>1) and (not word.replace('.','',1).isnumeric()) \
                  and (not word.replace("'",'',2).isnumeric())]


    tagged_words = pos_tag(filtered_terms, lang='eng')

    stemmer = SnowballStemmer("english")
    wn_tags = {'N':wn.NOUN, 'J':wn.ADJ, 'V':wn.VERB, 'R':wn.ADV}
    wnl = WordNetLemmatizer()
    stemmed_tokens = []
    for tagged_token in tagged_words:
        term = tagged_token[0]
        pos  = tagged_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))
    return stemmed_tokens

def tokenizer(s):

    print("Tokenizer")
    tokens = word_tokenize(s)
    tokens = [word.replace(',','') for word in tokens ]
    tokens = [word for word in tokens if word.find('*')!=True and \
              word != "''" and word !="``" and word!='description' \
              and word !='dtype']
    return tokens

def preprocessor(s):
    s = s.lower()
    s = s.replace(',', '. ')
    print("preprocessor")
    return(s)
```

```
pd.set_option('max_colwidth', 32575)


df = pd.read_excel("D:\Work\Course Work\Semester 4\STAT 656\Lectures &
Assignment\Week 10\Week 10 Assignment\CaliforniaCabernet.xlsx")

# Setup constants
n_docs     = len(df['description'])
n_samples  = n_docs
m_features = None
s_words    = 'english'
ngram = (1,2)

# Setup reviews in list 'discussions'
discussions = []
for i in range(n_samples):
    discussions.append(("%s" %df['description'].iloc[i]))

cv = CountVectorizer(max_df=0.95, min_df=2, max_features=m_features,\
                     analyzer=analyzer_func, ngram_range=ngram)
tf = cv.fit_transform(discussions)

print("\nVectorizer Parameters\n", cv, "\n")


n_topics        = 9
max_iter        =  5
learning_offset = 20.
learning_method = 'online'

tf_idf = TfidfTransformer()
print("\nTF-IDF Parameters\n", tf_idf.get_params(),"\n")
tf_idf = tf_idf.fit_transform(tf)


# Construct the IDF/TF matrix from the data
tfidf_vect = TfidfVectorizer(max_df=0.95, min_df=2, max_features=m_features,\
                             analyzer=analyzer_func, ngram_range=ngram)
tf_idf = tfidf_vect.fit_transform(discussions)
print("\nTF_IDF Vectorizer Parameters\n", tfidf_vect, "\n")

lda = LatentDirichletAllocation(n_components=n_topics, max_iter=max_iter,\
                                learning_method=learning_method, \
                                learning_offset=learning_offset, \
                                random_state=12345)
lda.fit_transform(tf_idf)
print('{:.<22s}{:>6d}'.format("Number of Reviews", tf.shape[0]))
print('{:.<22s}{:>6d}'.format("Number of Terms",     tf.shape[1]))
print("\nTopics Identified using LDA with TF_IDF")
tf_features = cv.get_feature_names()
max_words = 15
topic_description=[]
for index, topic in enumerate(lda.components_):
        message = "Topic #%d: " % index
        message += " ".join([tf_features[i]
                             for i in topic.argsort()[:-max_words - 1:-1]])
        topic_description.append(message[10:])
        print(message)
        print()

for i in range(len(topic_description)):
    topic_description[i]=topic_description[i].split(' ')

temp=lda.transform(tf_idf)
temp1=[]
for i in range(len(temp)):
    temp1.append(temp[i].argmax())
temp1=pd.DataFrame(temp1,columns=['Topic#'])
df=df.join(temp1)



table1=df.pivot_table(['points','price'],index='Topic#')
table1=table1.join(pd.DataFrame(topic_description))

table1=table1.rename_axis({'points':'avg_points','price':'avg_price'},axis=1)


table2=df.pivot_table('Review',index='Region',columns='Topic#',\
                      aggfunc='count',\
                      fill_value=0,margins=True)

def percentage_convert(x):
    for index in x.index:
        for i in x.columns:
            x.loc[index,i]=round(x.loc[index,i]*100/x.loc[index,'All'],2)

    return x
percentage_convert(table2)

print(table1.T)   #transposed table 1
print(table2)


#Export output to excel
with pd.ExcelWriter('D:\Work\Course Work\Semester 4\STAT 656\Lectures &
Assignment\Week 10\Week 10 Assignment\output.xlsx') as output:
    table1.T.to_excel(output,sheet_name='t1')
```

```
table2.to_excel(output,sheet_name='t2')
```

2) Results

	0	1	2	3	4	5	6	7	8
avg_points	90.08059939	84.5	87.26315789	90	85.64788732	89.07226107	82.09090909	84.5	86
avg_price	64.78789606	28.42857143	46.78947368	65	33.71529412	57.30023641	24	47	33.77777778
0	wine	barely	meet	punch	flavor	palate	sirah	brightness	bouquet
1	flavor	wait	coconut	expansive	blackberry	petit	petite	weedy	effort
2	tannin	sweaty	tightly	cardamom	cherry	verdot	cherry-berry	muscular	santa
3	black	bay	party	coast	dry	nose	bottling	breadth	light-bodied
4	blackberry	overpower	wound	aromatics	soft	merlot	showy	recall	elevation
5	cabernet	weave	lend	boast	drink	malbec	reduce	farm	lurk
6	currant	chile	fade	handful	wine	small	figure	opposite	loam
7	oak	front	saddle	enjoyment	sweet	franc	appropriately	cake	slate
8	year	tongue	beneath	central	oak	amount	curiously	black-fruit	ting
9	fruit	create	easygoing	tomato	cabernet	blend	provenance	relieve	notion
10	cherry	funky	pleasantly	amidst	finish	leather	dark-fruit	neighbor	excite
11	dry	drop	well-balanced	waft	tannin	juicy	root	lohr	gamy
12	rich	generosity	subdue	thickness	good	tar	awash	j.	offset
13	show	acceptable	bread	cracker	cab	pepper	lightness	six-plus	medium-weight
14	ripe	underbelly	small-production	graham	ripe	tobacco	pipe	today	reduction

Region	0	1	2	3	4	5	6	7	8	All
California Other	26.77	0	0	0	71.22	1.34	0.27	0	0.4	100
Central Coast	50.7	0.17	0.28	0	43.62	5	0	0	0.22	100
Central Valley	33.99	0.99	0.99	0	61.58	2.46	0	0	0	100
Clear Lake	0	0	0	0	100	0	0	0	0	100
High Valley	0	0	0	0	100	0	0	0	0	100
Lake County	50	0	0	0	50	0	0	0	0	100
Mendocino	60	0	3.33	0	36.67	0	0	0	0	100
Mendocino County	62.07	0	0	0	34.48	3.45	0	0	0	100
Mendocino Ridge	66.67	0	0	0	33.33	0	0	0	0	100
Mendocino/Lake Counties	56.12	0	0.51	0	42.86	0.51	0	0	0	100
Napa	78.31	0	0.14	0.03	18.43	2.98	0.04	0.05	0.03	100
Napa-Sonoma	70.24	0	0	0	21.43	8.33	0	0	0	100
North Coast	36.07	1.09	0	0	58.47	4.37	0	0	0	100
Red Hills Lake County	64.86	0	0	0	35.14	0	0	0	0	100
Redwood Valley	66.67	0	0	0	33.33	0	0	0	0	100
Sierra Foothills	48.41	0	0	0	45.24	5.56	0.79	0	0	100
Sonoma	65.22	0.31	0	0	30.79	3.29	0.22	0.18	0	100
South Coast	42.31	0	0	0	44.23	13.46	0	0	0	100
All	67.07	0.11	0.14	0.02	29.19	3.27	0.08	0.06	0.07	100