Applied Analytics

Text Analytics & Topic Analysis

Terminology
History
Linguistics
Statistical Machine Learning

STAT 656 – Applied Analytics using SAS© Enterprise Miner™

Week 8

Week 8

✔ Topics – Week 8

WEEK	DATE	TOPICS	ASSIGNMENT (due following week)
1	May 21	Introduction to Data Mining, Python and SAS Enterprise Miner	HW1: RIE & Linear Reg with Diamond Price Data
2	May 28	Data Preprocessing & Linear Regression	HW2: Linear Reg with Oil Production Data & 70/30 Assessment
3	June 4	Logistic Regression, The Confusion Matrix and Model Metrics	HW3: Logistic Regression with Fraud Detection
4	June 11	Decision Trees & Cross Validation	HW4: Cell Phone Activity Classification
5	June 18	Random Forests	Midterm Exam – Take Home
6	June 25	Neural Networks	HW6: Optional: Apply Keras to building a Neural Network
7	July 2	Genetic Algorithms for Advanced Feature Selection	HW7: Optional: Apply GA Selection to Diamond Prices
8	July 9	Speaker (6-7) & Introduction to Text Analytics (7-9)	HW8: Analysis of Wine Reviews in SAS EM and Python
9	July 16	Capstone Updates (6-7) & Topic Analysis (7-9)	HW9: Topic Analysis
10	July 23	Degree Plans (6-6:30) & Sentiment Analysis (6:30-9)	HW10: Sentiment Analysis
11	July 30	Final Exam & Web Scraping	Final Exam Open – Return Thursday, Aug 6
12	Aug. 10	Review of Final Exam & Grades	This is an optional Q&A Session

Text Analytics
Topic Analysis

✔ WEEK 8 LEARNING OBJECTIVES – Able to…

1. Understand unsupervised topic analysis in EM & Python
2. Practice Topic Analysis using NTHSA DATA

Applied Analytics

Week 7
Optional Assignment
GA Feature Selection

STAT 656 – Applied Analytics
using SAS© Enterprise Miner™

Week 7

✔ Optional Assignment – Feature Selection

Data: diamonds_train (Excel File) Select the best features then build a linear regression model using these features. Compare GA selection to Stepwise, the Full Model, and Lasso or Regularized Regression

Data Dictionary: DiamondsDictionary(PDF file) The target is an interval attribute labeled "price". The predictors consist of 6 interval features and 3 nominal features.

Hold-Out Sample: diamonds_validation(Excel file) File contains the same columns as the training data.

Assignment: This is an optional assignment, but since GA Selection is not in SAS EM, you can only use Python for this assignment.

Week 7

✔ Python Template for GA Features with Interval Target

Data: OilProduction(Excel File) Data on the oil production from fracking wells.

Data Dictionary: OilProduction_Data_Dictionary(PDF file) The target is an interval attribute labeled "Log_Cum_Production". The predictors consist of 10 interval and 2 nominal features.

Template: GA_OilProduction(Python .py file) This code contains the latest template for implementing GA Selection, Stepwise and Lasso Selection for Interval Targets

Assignment: Modify this template to conduct a similar analysis of the diamond price data.

Week 7

✔ Python Imports for GA Features with Interval Target

deap: A python package for genetic algorithms. Install using from inside your working Anaconda environment:

conda install -c conda-forge deap

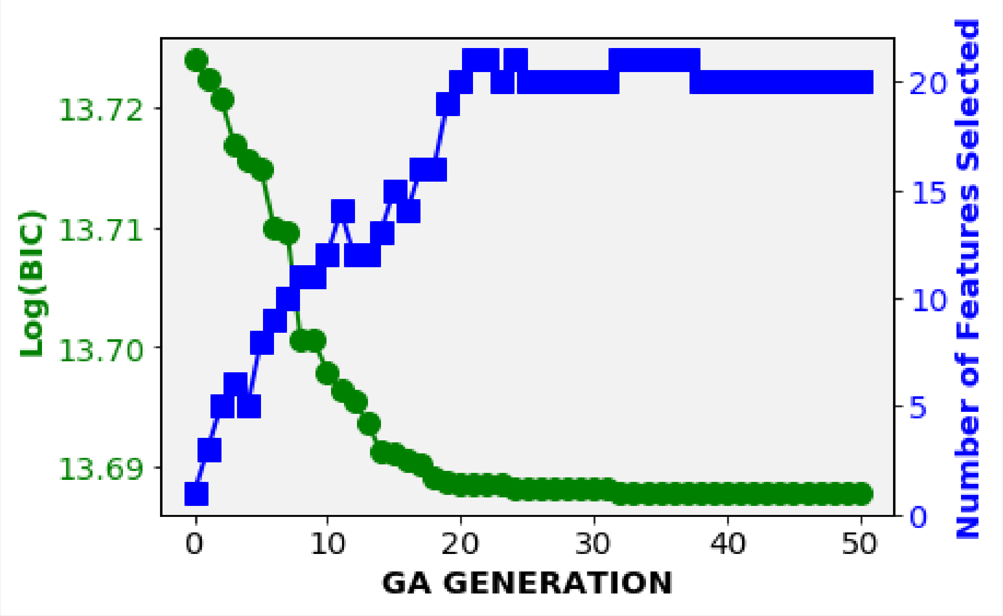
```
from deap import creator, base, tools, algorithms

import random, sys, time, warnings
import pandas as pd
import numpy  as np
import matplotlib.pyplot                  as plt
import statsmodels.api                    as sm
import statsmodels.tools.eval_measures as em
from AdvancedAnalytics.ReplaceImputeEncode import ReplaceImputeEncode, DT
from AdvancedAnalytics.Regression          import linreg, stepwise
from math                 import log, isfinite, sqrt, pi
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression, Lasso
from sklearn.metrics      import mean_squared_error, r2_score, log_loss
from scipy.linalg         import qr_multiply, solve_triangular
```

Week 7

✔ Solution

GA_HW7.py: A solution to this assignment is posted in the week 7 assignment folder.



GA Solution: Found you can minimize BIC with 20 features: carat, depth, table, x, and most of the levels from the 3 nominal features..

Week 7

✔ Python Template for GA Features with Interval Target

Full Model: 23 Features. Train BIC = 880,159 & Validation ASE = 471,688

Interval Features(6): carat, depth, table, x, y and z

Nominal Features(17): All except the last column for each nominal feature.

- Cut(5-1): All except cut=very good
- Color(7-1): All levels except color=J
- Clarity(8-1): All levels except clarity=VVS2

GA Selection: 20 Features. Train BIC = 880,133 & Validation ASE = 472,808

Interval Features(4): carat, depth, table and x

Nominal Features(16):

- Cut(5-2): Fair, Good and Ideal (missing 2)
- Color(7-1): All levels except color=G
- Clarity(8-1): All levels except clarity=VS1

Week 7

✔ All Features: logreg.display_split_metrics

```
***************************************************************
**************      FIT FULL MODEL      **************
***************************************************************


Model Metrics..........        Training      Validation
Observations...........           51999            1941
Coefficients...........              24              24
DF Error...............           51975            1917
R-Squared..............          0.9204         -0.5395
Adj. R-Squared.........          0.9203         -0.5580
Mean Absolute Error....        752.6686        584.1012
Median Absolute Error..        541.2481        536.8049
Avg Squared Error......    1307150.7510     471687.6517
Square Root ASE........       1143.3069        686.7952
Log Likelihood.........    -439943.7116    -15432.8419
AIC           .........     879937.4232     30915.6837
AICc          .........     879937.4482     30916.3626
BIC           .........     880158.8977     31054.9577
```

Week 7

GA Selection from 50 Generations

```
**********************************************************************
**************     GA Selection using   bic Fitness         ***********
************** statsmodels Models and   star Initialization ***********

gen     nevals  features        range   min     avg     max             Ln(Fit)
0       27      1               98883   912602  999316  1.01149e+06     13.7241
1       26      3               100326  911156  949506  1.01148e+06     13.7225
2       22      5               30560.9 909707  919459  940268          13.7209
3       21      6               6391.54 906211  910627  912602          13.717
4       25      5               4705.73 905002  907761  909707          13.7157
5       26      8               4741.6  904351  905874  909093          13.715
6       27      9               6557.05 899844  904612  906401          13.71
7       23      10              5463.23 899535  901566  904998          13.7096
8       23      11              33574.1 891437  900772  925011          13.7006
9       25      11              8296.67 891437  898081  899734          13.7006
10      25      12              10533   889002  894046  899535          13.6979
11      24      14              7616.86 887704  890498  895321          13.6964
12      24      12              3822.19 886963  888837  890785          13.6956
13      24      12              10872.6 885311  888193  896183          13.6937
14      22      13              12692.2 883149  887533  895841          13.6912
15      26      15              4013.76 882949  885091  886963          13.691
16      20      14              35537.7 882607  884652  918145          13.6906
17      26      16              4761.98 882266  883016  887028          13.6902
```

```
49      26      20              3354.25 880133  880488  883487          13.6878
50      25      20              6.268   880133  880133  880139          13.6878
GA Runtime  50.29810309410095  sec.
Individuals in HoF:  306
Best Fitness: 880133.0512437393
Number of Features Selecterd:  20

Features: ['carat', 'depth', 'table', 'x', 'cut0:Fair', 'cut1:Good', 'cut2:Ideal', 'color0:D', 'color1:E', 'color2:F',
'color4:H', 'color5:I', 'color6:J', 'clarity0:I1', 'clarity1:IF', 'clarity2:SI1', 'clarity3:SI2', 'clarity5:VS2',
'clarity6:VVS1', 'clarity7:VVS2']
```

Week 7

✔ GA Selection

OLS Regression Results

Dep. Variable:	price	R-squared:	0.920
Model:	OLS	Adj. R-squared:	0.920
Method:	Least Squares	F-statistic:	3.003e+04
Date:	Wed, 08 Jul 2020	Prob (F-statistic):	0.00
Time:	16:30:41	Log-Likelihood:	-4.3995e+05
No. Observations:	51999	AIC:	8.799e+05
Df Residuals:	51978	BIC:	8.801e+05
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7081.4814	390.716	18.124	0.000	6315.675	7847.288
carat	1.115e+04	49.869	223.612	0.000	1.11e+04	1.12e+04
depth	-66.9923	4.228	-15.844	0.000	-75.280	-58.705
table	-27.0153	3.002	-8.999	0.000	-32.899	-21.131
x	-980.7739	21.087	-46.511	0.000	-1022.105	-939.443
cut0:Fair	-754.0449	32.470	-23.222	0.000	-817.687	-690.403
cut1:Good	-169.1948	18.589	-9.102	0.000	-205.630	-132.760
cut2:Ideal	82.6635	12.880	6.418	0.000	57.419	107.908
color0:D	472.4248	18.254	25.881	0.000	436.647	508.202
color1:E	271.7293	16.332	16.638	0.000	239.718	303.740
color2:F	211.4119	16.272	12.993	0.000	179.519	243.305
color4:H	-503.8405	16.922	-29.774	0.000	-537.008	-470.673
color5:I	-986.1247	19.583	-50.356	0.000	-1024.507	-947.742
color6:J	-1912.3640	24.825	-77.033	0.000	-1961.022	-1863.707
clarity0:I1	-4641.1988	46.047	-100.794	0.000	-4731.450	-4550.947
clarity1:IF	772.3894	30.557	25.277	0.000	712.497	832.282
clarity2:SI1	-914.8049	16.658	-54.916	0.000	-947.455	-882.154
clarity3:SI2	-1898.8833	18.350	-103.484	0.000	-1934.849	-1862.918
clarity5:VS2	-306.9829	16.624	-18.466	0.000	-339.566	-274.400
clarity6:VVS1	432.7541	23.438	18.464	0.000	386.816	478.692
clarity7:VVS2	384.6304	20.956	18.354	0.000	343.557	425.704

Week 7

GA Selection: logreg.display_split_metrics

```
Training Data Metrics
ASE.....................1307319.9522
Square Root of ASE.......    1143.3809
AIC..................... 879938.1537
BIC..................... 880133.0512
Adj. R-Squared..........       0.9203

Validation Data Metrics
ASE..................... 472808.3070
Square Root of ASE.......     687.6106


Model Metrics..........          Training       Validation
Observations...........             51999             1941
Coefficients...........                21               21
DF Error...............             51978             1920
R-Squared..............            0.9203          -0.5432
Adj. R-Squared.........            0.9203          -0.5593
Mean Absolute Error....          753.0048         585.1926
Median Absolute Error..          541.8592         535.7668
Avg Squared Error......      1307319.9522      472808.3070
Square Root ASE........         1143.3809         687.6106
Log Likelihood.........      -439947.0768     -15435.1449
AIC           .........       879938.1537      30914.2897
AICc          .........       879938.1732      30914.8174
BIC           .........       880133.0512      31036.8508
```

Week 7

✔ Stepwise vs. Lasso (same features different model)

Stepwise: 21 Features. Train BIC = 880,139 & Validation ASE = 471,775

Interval Features(4): carat, depth, table and x

Nominal Features(17): All except the last column for each nominal feature.

- Cut(5-1): All except cut=very good
- Color(7-1): All levels except color=J
- Clarity(8-1): All levels except clarity=VVS2

Lasso: 21 Features. Train BIC = 880,945 & Validation ASE =458,401

Interval Features(4): carat, depth, table and x

Nominal Features(17): missing last column of each nominal feature

- Cut(5-1): Fair, Good, Ideal and Good
- Color(7-1): All levels except color=J
- Clarity(8-1): All levels except clarity=VVS2

Week 7

✔ stepwise: logreg.display_split_metrics

```
Model Metrics..........         Training     Validation
Observations...........            51999           1941
Coefficients...........               22             22
DF Error...............            51977           1919
R-Squared..............           0.9203        -0.5398
Adj. R-Squared.........           0.9203        -0.5567
Mean Absolute Error....         752.6798       584.1389
Median Absolute Error..         541.1422       536.8332
Avg Squared Error......     1307204.5307    471775.4825
Square Root ASE........        1143.3305       686.8591
Log Likelihood.........     -439944.7813    -15433.0225
AIC        ............      879935.5626     30912.0451
AICc       ............      879935.5838     30912.6210
BIC        ............      880139.3191     31040.1771
```

Week 7

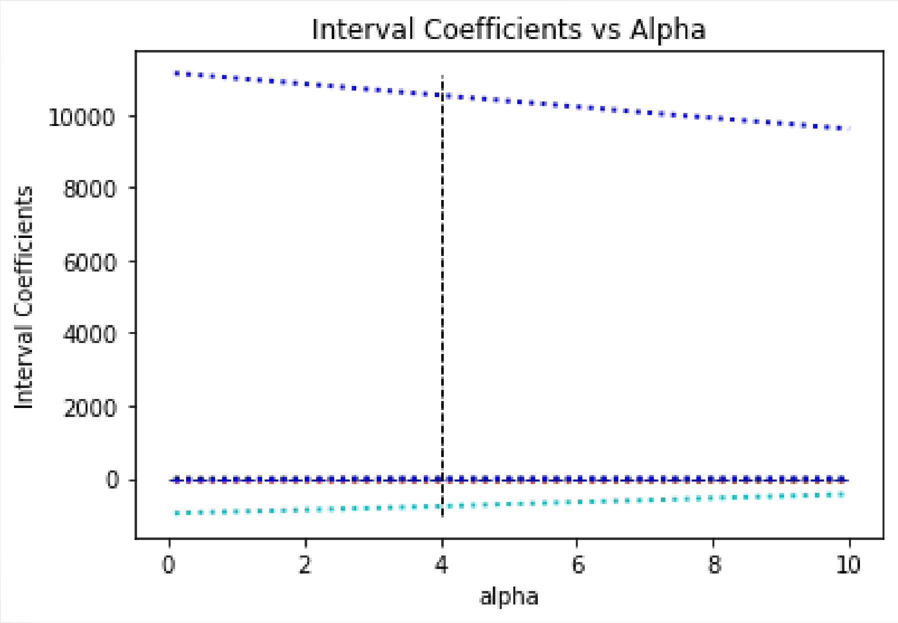
Lasso alpha=4: logreg.display_split_metrics

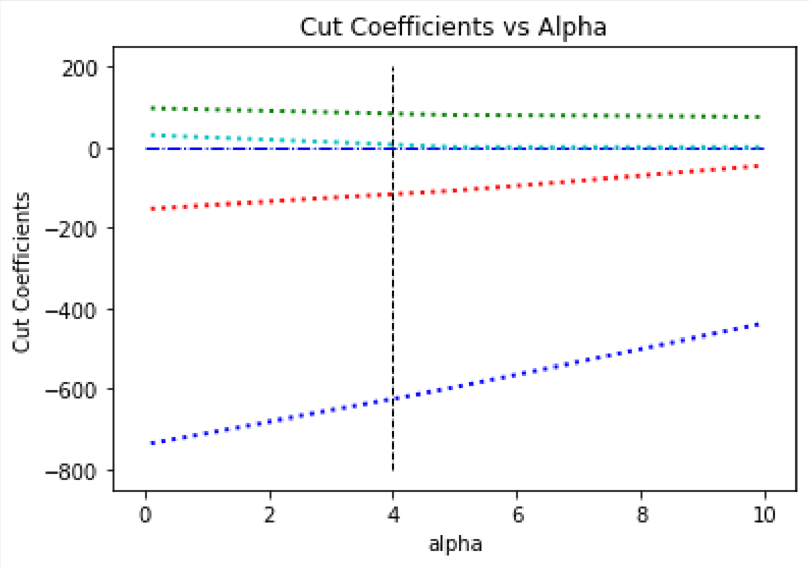
```
Alpha:   4.0  Number of Coefficients:  21 / 23


Model Metrics..........        Training    Validation
Observations...........           51999          1941
Coefficients...........              22            22
DF Error...............           51977          1919
R-Squared..............          0.9191       -0.4962
Adj. R-Squared.........          0.9191       -0.5126
Mean Absolute Error....        751.9146      581.4569
Median Absolute Error..        529.6072      551.8614
Avg Squared Error......    1327617.8264   458401.9667
Square Root ASE........       1152.2230      677.0539
Log Likelihood.........    -440347.6518  -15405.1141
AIC          ..........     880741.3036   30856.2283
AICc         ..........     880741.3248   30856.8042
BIC          ..........     880945.0601   30984.3603
```

Week 7

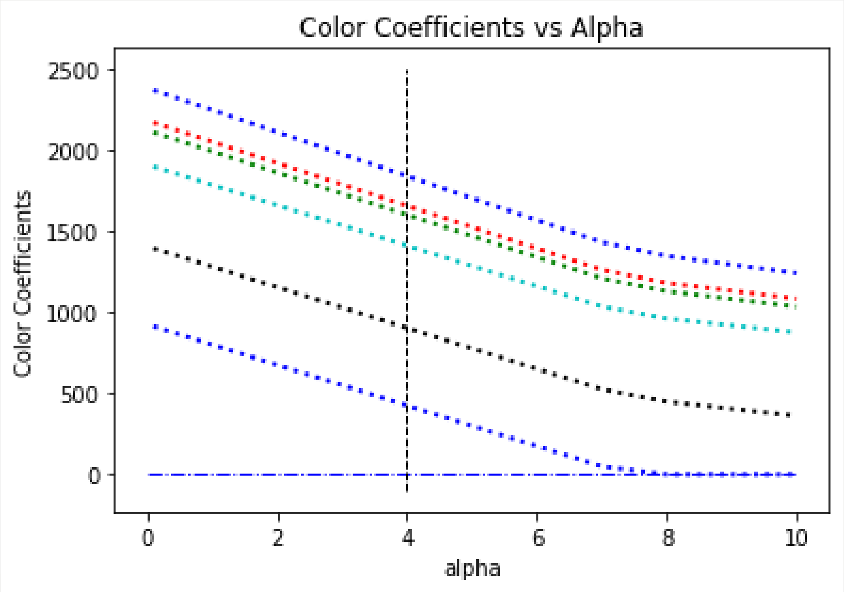
✔ Lasso Coefficients vs Alpha

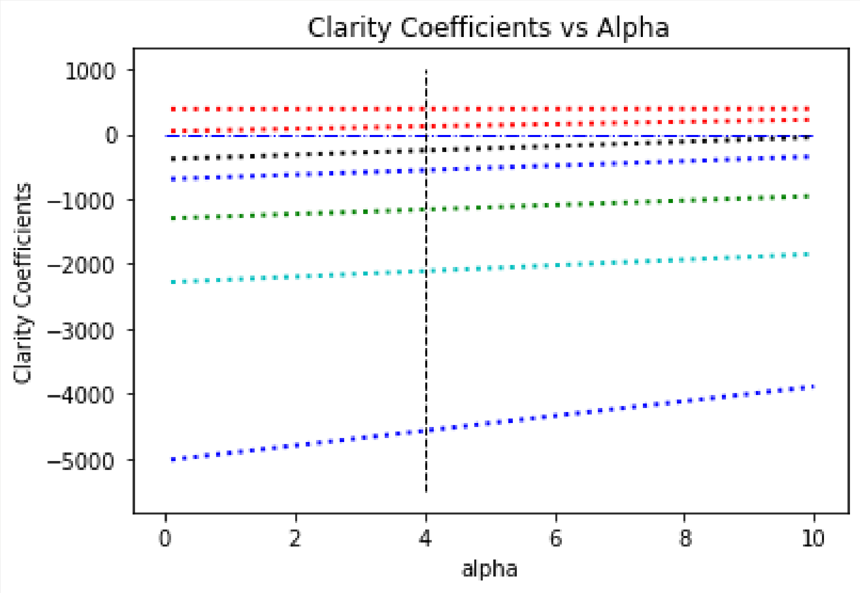




Week 7

Lasso Coefficients vs Alpha





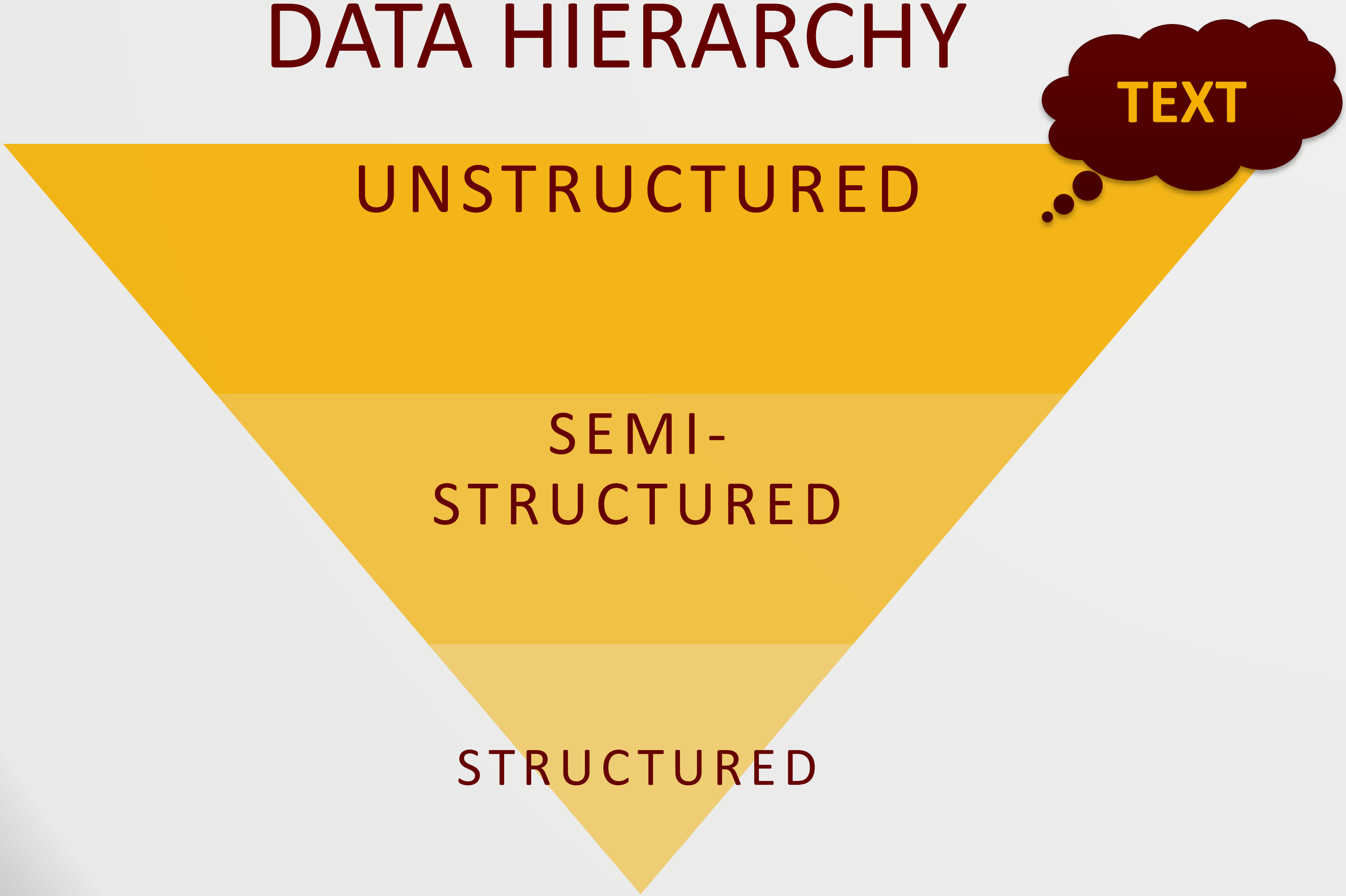
Applied Analytics

Introduction to Text Analytics

Week 8

STAT 656 – Applied Analytics using SAS© Enterprise Miner™

DATA HIERARCHY



Linguistics

History & Evolution 1950-1985

Rationalist Views

- Advocated by Noam Chomsky (MIT)

  "Language is an innate human faculty. Children do not learn natural language from limited input during their early years"

- Humans are born with language processing, not just pattern recognition.
- AI researchers developed application-specific rules and algorithms for solving problems.

Linguistics

History & Evolution Post 1985

Rise of Empiricist & Computational Linguistics

- Computers and large collections of annotated text became available.
- Humans are born with the biology for learning language, not a detailed set of language rules.
- Language is learned using pattern recognition and statistical thinking.
- Example:

  "The British left waffle on the Falklands."

Corpus: Linguistic Data (60s-Present)

- Brown Corpus (1960s)
  - Developed at Brown University by Kucera and Francis
  - Balanced annotated corpus of 500 fiction and nonfiction documents
  - Used for IR Research – Statistical Similarity of Documents
- London-Lund Corpus (LLC, 1970s)
- Lancaster-Oslo-Bergen (LOB, 1980s) Corpus
  - Similar to Brown Corpus, but for British English
- Penn TreeBank Corpus (1980s)
  - Developed at Penn State University
  - 2,499 Wall Street Journal stories.
  - Available in different languages, and as recorded speech.
- American National Corpus (ANC, 2000s)
  - 22 million word subcorpus
- Google N-Gram Corpus (2000s) (up to 5-gram)
  - 1 trillion words from web pages
- New ISO and Specialized Corpus Developed (2010s)
  - Twitter, Facebook and Blogs
  - Specialized Domains

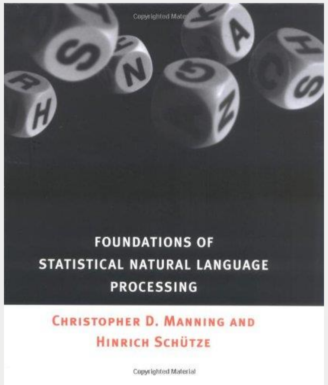
Computational Linguistics

- The use of rule-based and/or statistical modeling of natural language
- Rule-based approaches are rules for extract parts-of-speech and convert raw text into meaningful data structures.
  - Used for speech recognition and synthesis
- Statistical approaches use annotated text, dictionaries and rules to model the relationships between text and meaning.
  - Used for POS tagging, colocation, entity Extraction
  - Reduce language ambiguity
  - Used for extracting domain specific knowledge

Statistical Machine Learning

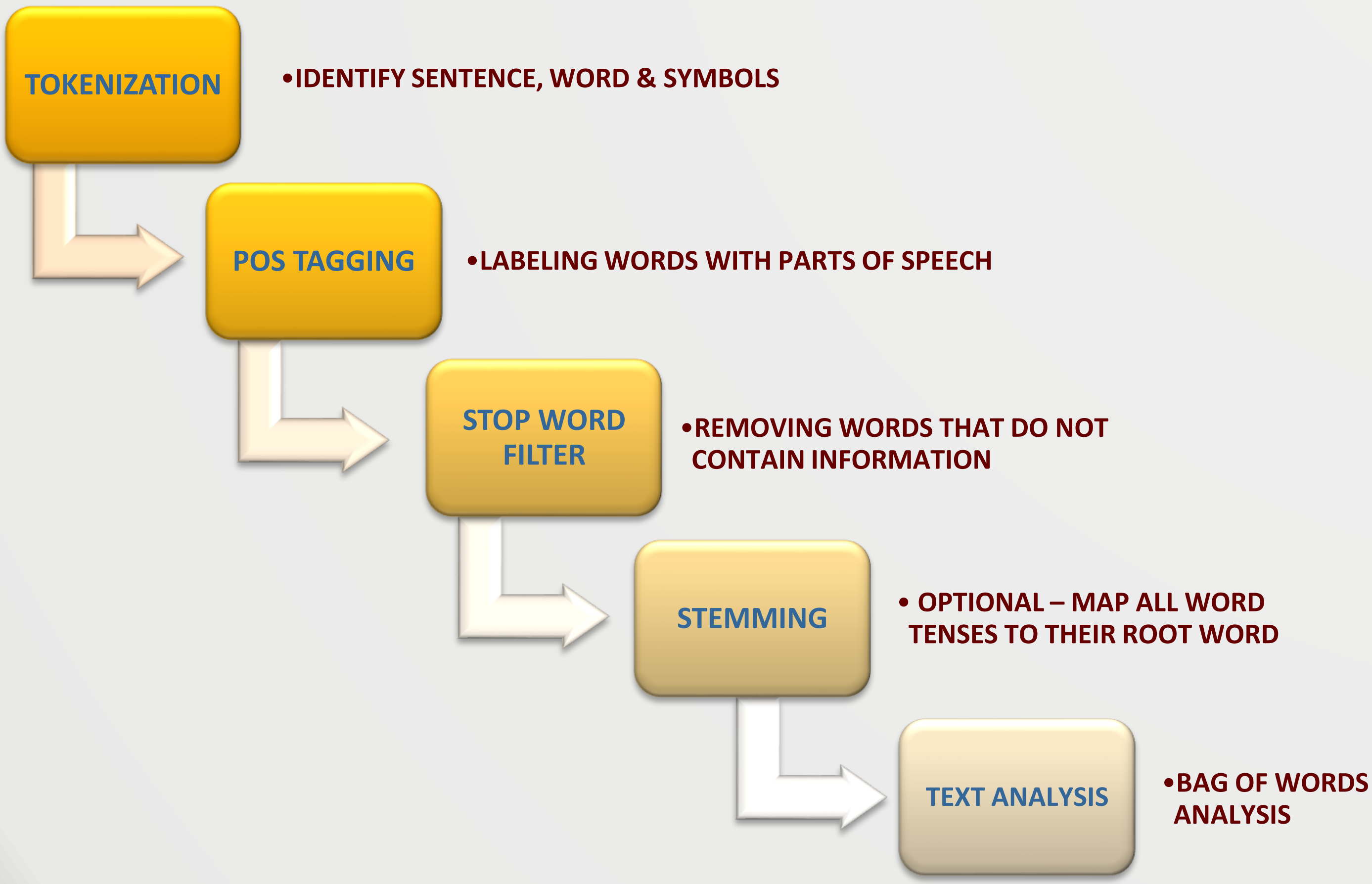
- Statistical approaches for NLP that develop models from corpus designed to discover topics.
- The annotated corpus is data where each word is tagged with a POS.
- Statistical Machine Learning develops a model for predicting word POS, identifying noun groups, entities and topics.
- A corpus is valuable for customizing the model. A corpus is language specific and often topic specific.

Linguistics – Statistical Machine Learning



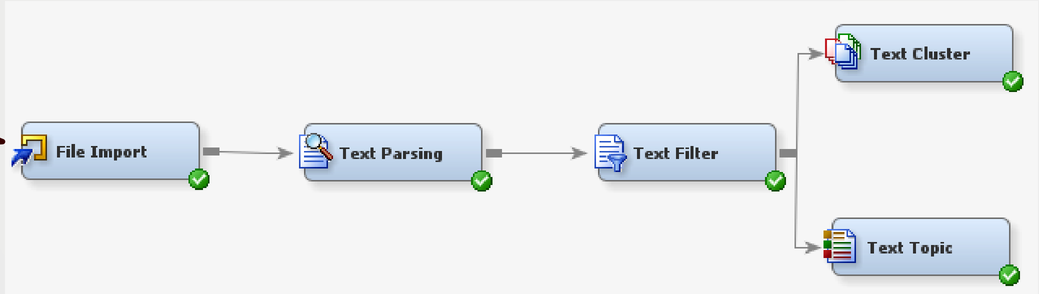
- Manning C.D. and Schutze, H. (1999) Foundations of Statistical Natural Language Processing, MIT Press.
- Indurkhya, N. and Damerau, R. J. (2010) Handbook of Natural Language Processing; CRC Press.
- Pustejovsky, J. and Stubbs, A. (2012) Natural Language Annotation for Machine Learning; O'Reilly.
- Others – Google Stanford Natural Language Laboratory

Text Analysis using Word Counts

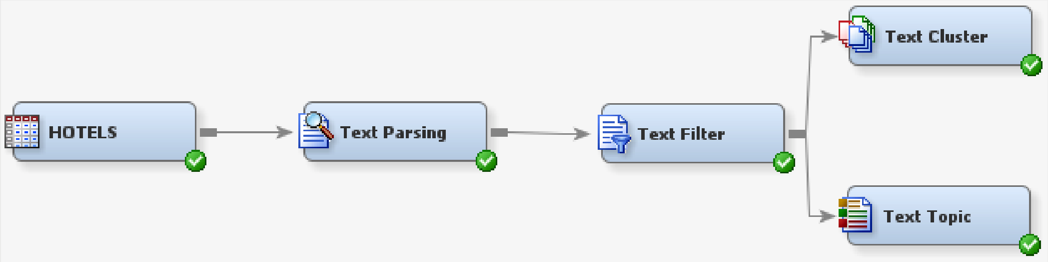


✔ Approaches to Text Analysis in Non-HP SAS Enterprise Miner: Sample-Parse-Filter-Cluster

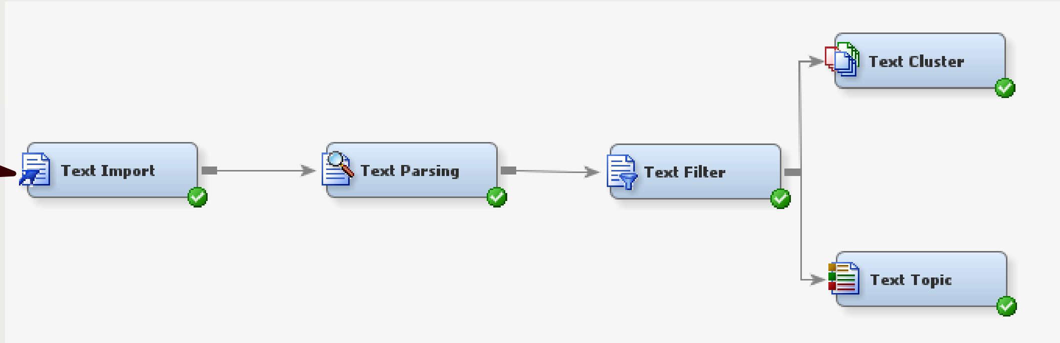
Import Single File (Excel) with Column of Text Entries



Read Single File (SAS) with a Column of Text Entries

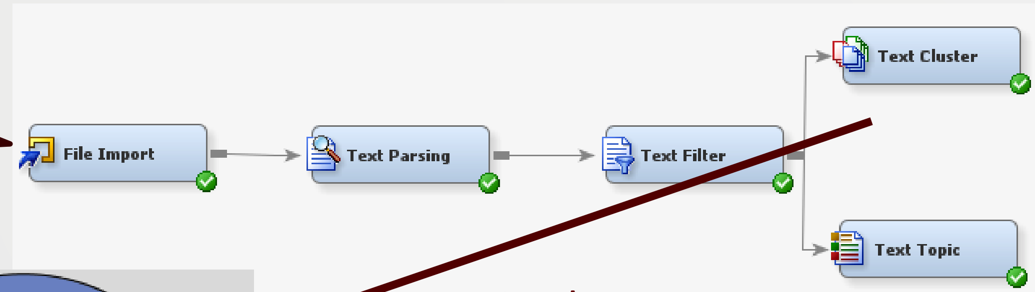


Import Multiple Files from a Single Directory

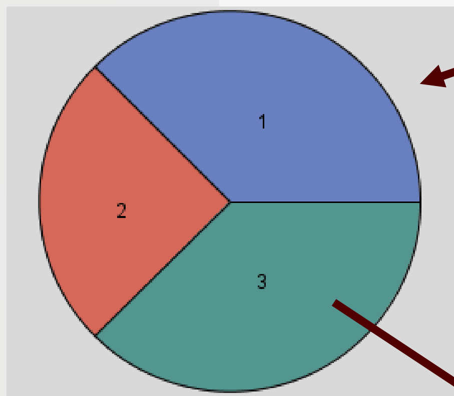


Approaches to Text Analysis in Non-HP SAS Enterprise Miner: Sample-Parse-Filter-Cluster

Import Single File (Excel) with Column of Text Entries



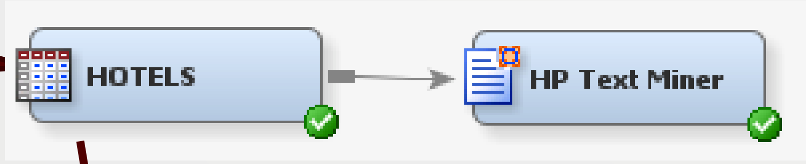
Topic Clusters



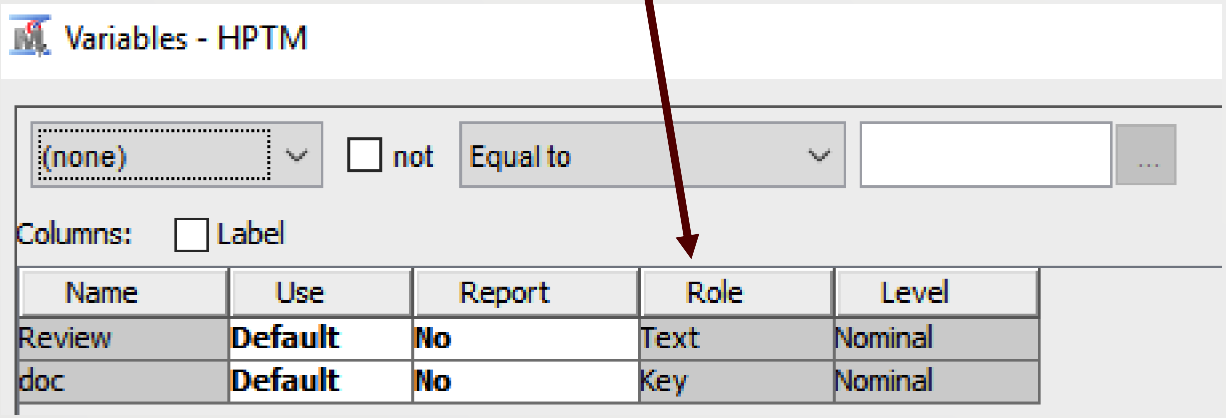
Cluster ID	Descriptive Terms
1	pale past white felt real running sound left going half forward air close best dead ...
2	fat salt falls daily cool north air escape fall fast great passing light short bent ...
3	home fair round fine bad quiet general right light down good passing black last part ...

Approaches to Text Analysis in HP SAS Enterprise Miner: Sample-Parse-Filter-Cluster

Read Single File (SAS) or Import a File (Excel) with a Column of Text Entries and a Key Column.



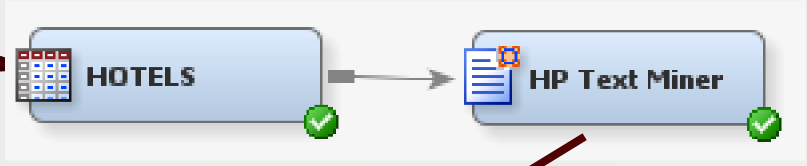
Set Variable Roles



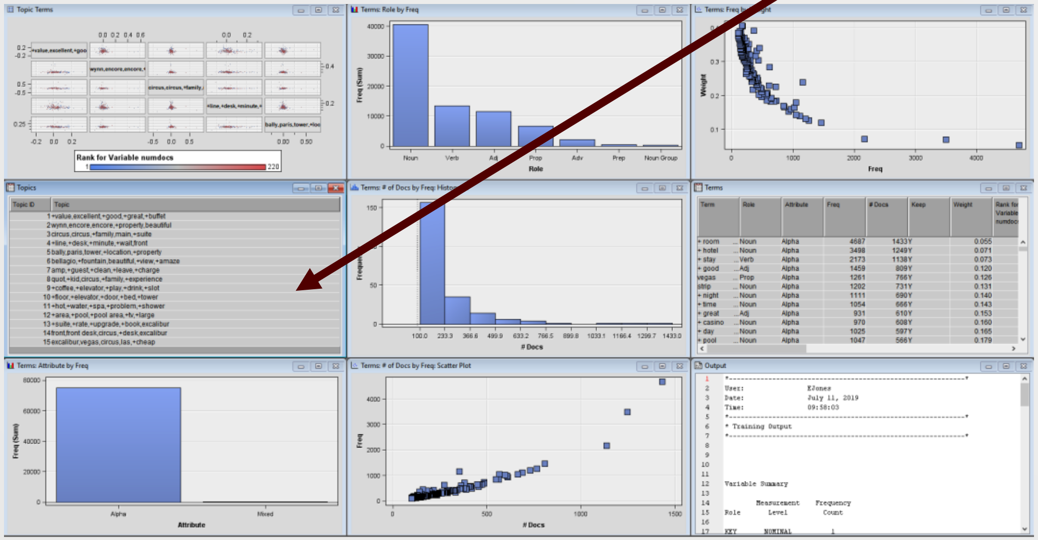
Name	Use	Report	Role	Level
Review	**Default**	**No**	Text	Nominal
doc	**Default**	**No**	Key	Nominal

Approaches to Text Analysis in HP SAS Enterprise Miner: Sample-Parse-Filter-Cluster

Read Single File (SAS) or Import a File (Excel) with a Column of Text Entries and a Key Column.



Topic Clusters



Term Frequency

- The Term Matrix is a t by d matrix where t is the number of terms (words) in the collection and d is the number of documents.
- Element f_{ij} is the number of times the ith term occurs in the jth document.
- Many of the term frequencies are zero.
- Raw term frequencies are larger for larger documents.

Term-Doc Matrix

✔ A matrix with t rows and d columns

DOCUMENTS

TERMS

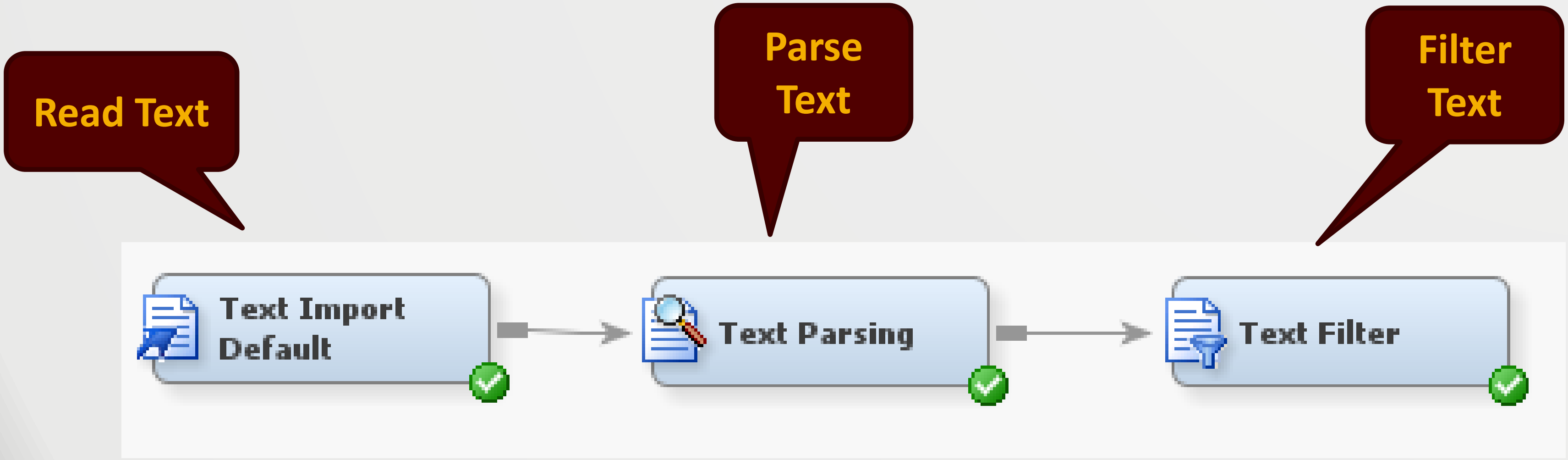
$$
\begin{Bmatrix}
f_{11} & f_{12} & f_{13} & \dots & f_{1d} \\
f_{21} & f_{22} & f_{23} & \dots & f_{2d} \\
f_{21} & f_{22} & f_{33} & \dots & f_{3d} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
f_{t1} & f_{t2} & f_{t3} & \dots & f_{td}
\end{Bmatrix}
$$

Term Frequencies by Document

- Many values will be zero.
- Larger docs will have larger term frequencies

Term-Doc Matrix

✔ Create the Matrix – Sample-Parse-Filter



Term-Doc Matrix

✔ Import

Usually Increase Maximum Text Size Above 100

Consider Using Synonym List

Consider Changing Frequency & Term Weights

Default has no Synonyms

Text Filter Node

✔ Text Filter Node – Frequency Weights

Frequency Weighting Methods

The following frequency weighting functions, g(.), are available in the **Text Filter** node.

- **Default**

  The default frequency weighting method is Log with one exception. In a process flow that has multiple **Text Filter** nodes, the default frequency weighting method that is used in a node is determined by the setting that was specified in the previous **Text Filter** node.

- **Binary**

  an indicator function, where $g(f_{ij})= 1$ if a term appears in the document, and $g(f_{ij})= 0$ if it does not. This function removes the effect of terms that occur repeatedly in the same document.

- **Log**

  $g(f_{ij})= \log_2(f_{ij}+1)$. This function dampens the effect of terms that occur many times in a document.

- **None**

  $g(f_{ij})= 1$. In other words, no change is applied to the raw frequency for the term.

Default Frequency Weight is Log(f+1)

Must Set to None for Sentiment Analysis

Text Filter Node

✔ Text Filter Node – Term Weights

- **Entropy**

$$w_i = 1 + \sum_j \frac{(f_{ij}/g_i)\log_2(f_{ij}/g_i)}{\log_2(n)}$$

Here, g_i is the number of times that term i appears in the document collection, and n is the number of documents in the collection. Log(.) is taken to be 0 if f_{ij}=0. This method gives greater weight to terms that occur infrequently in the document collection by using a derivative of the entropy measure found in information theory.

- **Inverse Document Frequency**

$$w_i = \log_2\left(\frac{1}{P(t_i)}\right) + 1$$

Here, $P(t_i)$ is the proportion of documents that contain term t_i. This method gives greater weight to terms that occur infrequently in the document collection by placing the number of documents that contain the term in the numerator of the formula.

- **Mutual Information**

$$w_i = \max_{C_k}\left[\log\left(\frac{P(t_i, C_k)}{P(t_i)\,P(C_k)}\right)\right]$$

Here, $P(t_i)$ is the proportion of documents that contain term t_i, $P(C_k)$ is the proportion of documents that belong to category C_k, and $P(t_i, C_k)$ is the proportion of documents that contain term t_i and belong to category C_k. Log(.) is taken to be 0 if $P(t_i, C_k)$=0 or $P(C_k)$=0.

This weight is valid only if the data source includes a categorical target variable. The weight is proportional to the similarity of the distribution of documents that contain the term to the distribution of documents that are contained in the respective category.

- **None**

w_i= 1. In other words, no term weight is applied.

Default Weight is Entropy for Interval Targets

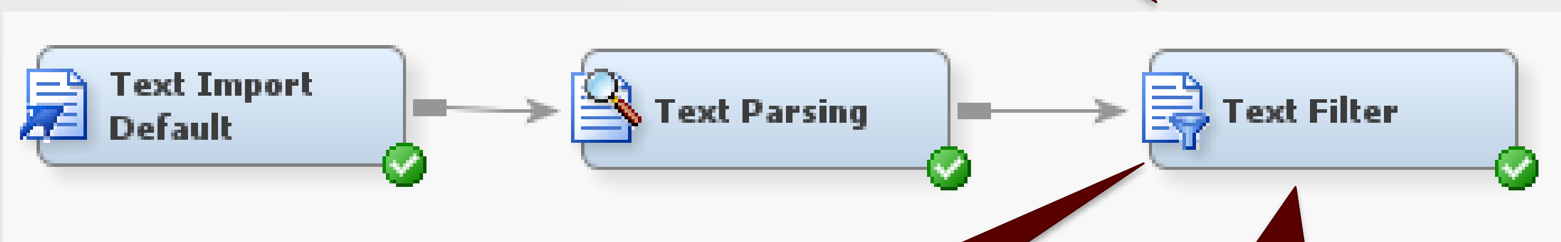
IDF – Popular Alternative

Default Weight is Mutual Information for Categorical Targets

Term-Doc Matrix

✔ Import

Default Frequency Weight is Log(f+1)



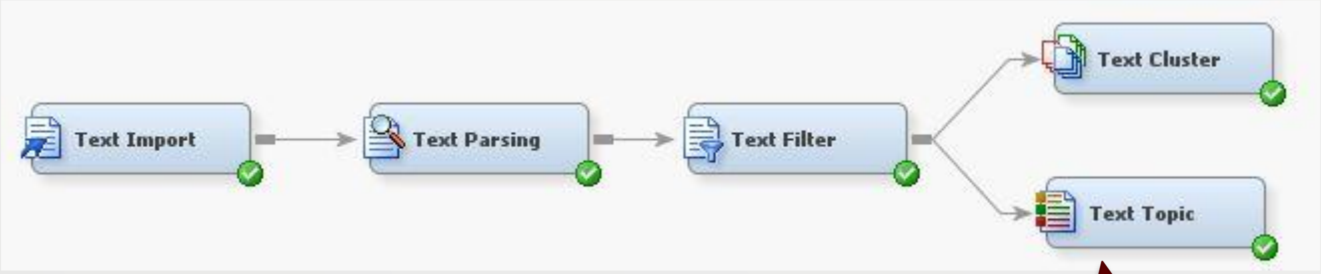
Default Term Weight is Entropy or Mutual Information

Consider Setting Term Weight to IDF

Topic Analysis

- ✔ Two Approaches

Cluster Analysis Node



Topic Analysis Node

Topic Analysis

✔ Two Approaches in SAS EM

- **Cluster Analysis**
  - **Mathematically Determined**
  - **Independent Topics**:
    Documents assigned only 1 topic
- **Topic Analysis Node**
  - **User Driven**
  - Allows for **Non-Independent Topics**:
    Documents can be assigned multiple topics

Topic Analysis

✔ SVD Based

- Both approaches start with a **Singular Value Decomposition** of the Term-Document Matrix
- The Matrix can be weighted and customized with POS, Synonyms, Stop Words and Stemming
- Cluster Analysis is the **SVD solution followed by** the **Min-Max Rotation** of the Solution to better identify independent document clusters.
- Topic Analysis uses SVD and Rotations but allows user to further refine topic clusters.

SVD

✔ Singular Value Decomposition of Term-Doc Matrix

DOCUMENTS

TERMS

$$
\begin{Bmatrix}
f_{11} & f_{12} & f_{13} & \cdots & f_{1d} \\
f_{21} & f_{22} & f_{23} & \cdots & f_{2d} \\
f_{21} & f_{22} & f_{33} & \cdots & f_{3d} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
f_{t1} & f_{t2} & f_{t3} & \cdots & f_{td}
\end{Bmatrix}
$$

Term-Document Matrix (T)

$$T = USV'$$

Identifies Term Clusters

SVD: Singular Value Decomposition

SVD Transforms the Term-Document Matrix - T:

$$T = USV'$$

Where U is a t x k **orthonormal** matrix, with t=number of terms
V is k x d **orthonormal**, with d=number of documents, and
Σ is k x k **diagonal** matrix containing the eigenvalues of T.
k is the rank of T, where k ≤ t and k ≤ d.

Normally k is equal to the number of topic clusters

SVD Rotations

The Term-Document matrix can be rotated:

$$T = U \Sigma V' = (U \cdot P)\Sigma(P \cdot V)' = U_* \Sigma V_*'$$

- Given an SVD solution there are an infinite number of equally good solutions defined by P.
- Most software do not allow for rotations. The SAS EM uses a *Varimax Rotation* to better identify meaningful topic clusters.
- This rotation is not done in Python SVD.

Normalized Term Frequency

- Term frequencies depend on the size of the document.
- There are several ways to normalize the term frequencies.
- Normalize using largest frequency or total word count
  - $tf_{ij} = f_{ij} \,/\, max(f_{1j}, f_{2j}, \ \ldots f_{nj})$
  - $tf_{ij} = f_{ij} \,/\, sum(f_{1j}, f_{2j}, \ \ldots f_{nj})$

Inverse Document Frequency

- If d = total number of documents, the idf_i for the ith term is:
  - $idf_i = log(d/d_i)$ *where* d_i *is the number of documents in the collection that contain the ith term.*
  - $0 \leq idf_i \leq log(d)$, *for all terms*
- The inverse document frequency is a measure of how many documents contain the ith term.
  - $idf_i = 0$, *if every document contains the ith term*
  - $idf_i = log(d)$, *if the ith term appears in only one document*

IDF for Python Sklearn

- If d = total number of documents, the idf_i for the ith term is:
  - $idf_i = log[(d+1)/(d_i+1)]$ where d_i=number of documents containing the ith term.
  - $0 \leq idf_i \leq log((d+1)/2)$, for all terms
- The inverse document frequency is a measure of how many documents contain the ith term.
  - $idf_i = 0$, if every document contains the ith term
  - $idf_i = log[(d+1)/2]$, if the ith term appears in only one document

TF-IDF (most software)

- The term weight TF-IDF is the term frequency weighted by the inverse document frequency.
  - $w_{ij} = f_{ij} x \; idf_i$
  - *If all documents contain the ith term,* $w_{ij} = 0$ *.*
  - *If only one document contains the ith term then the raw term frequency* f_{ij} *is increased by log(d).*
  - *Essentially TF-IDF reduces the term frequency for terms that appear in all or most documents and increases the term frequency when the term appears only in a small number of documents.*

TF-IDF (Python sklearn)

- The term weight TF-IDF is the term frequency weighted by the inverse document frequency.
  - $w_{ij} = f_{ij} x \log[(d+1)/(d_i +1)]$
    If all documents contain the ith term, $w_{ij} = 0$.
  - *If only one document contains the ith term then the raw term frequency* f_{ij} *is increased by log(d/2).*
  - *Essentially TF-IDF reduces the term frequency for terms that appear in all or most documents and increases the term frequency when the term appears only in a small number of documents.*

Other Term Weights

- TF-IDF is only one of a class of term weighting schemes. In general the weight term weight can be expressed as:
  - *$t_{ij} = wgt_i \times g(f_{ij})$, where $g(f)$ is a non-decreasing function of f.*
- Typical values for the term weight *wgt_i* are:
  - *Entropy*
  - *Mutual Information for applications with a categorical target*
  - *$wgt_i = 1$*

Choices for *g(f)*

- The function $g(f_{ij})$ *is typically selected to reduce the extreme spread in term frequencies. The most frequent terms often have values over 1,000; whereas the 10th ranked term might have a value in the 10's.*
- Typical values for the term function are:
  - $g(f_{ij}) = 1$
  - $g(f_{ij}) = \log(f_{ij}+1)$
  - Binary: $g(f_{ij}) = 1$ if $f_{ij} > 0$; otherwise $g(f_{ij}) = 0$

Recommendations

- The choice for wgt_i and $g(f_{ij})$ depends on the application.

- Applications where the documents in the collection are about the same size and the documents are large, the choice for wgt_i and $g(f_{ij})$ are not important.

- If the raw term frequencies are large for a few terms and smaller for the rest, use $g(f_{ij}) = log(f_{ij}+1)$ or the binary function.

- If the size of the documents vary considerably, weight $g(f_{ij})$ by the inverse document frequency $wgt_i = log(d/df_i)$

Applied Analytics

Text Cluster Node

STAT 656 – Applied Analytics using SAS© Enterprise Miner™

Topic Analysis

✔ Customer Reviews of Las Vegas Hotels

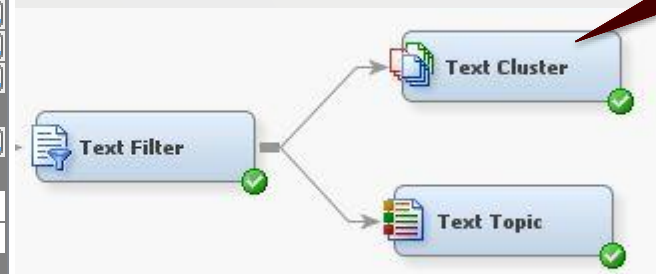
- **Data**
  - 1,671 customer reviews of Las Vegas Hotels
- **Objective**
  - Identify major review topics (categories)
  - Identify hotels associated with topics

Cluster Analysis

✔ Cluster Analysis Default Properties

Property	Value
General	
Node ID	TextCluster
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	7/20/16 2:42 PM
Run ID	51f8697e-6882-49cd-b3e5-e41
Last Error	
Last Status	Complete
Last Run Time	7/20/16 8:56 PM
Run Duration	0 Hr. 0 Min. 7.21 Sec.
Grid Host	
User-Added Node	No

Cluster Analysis



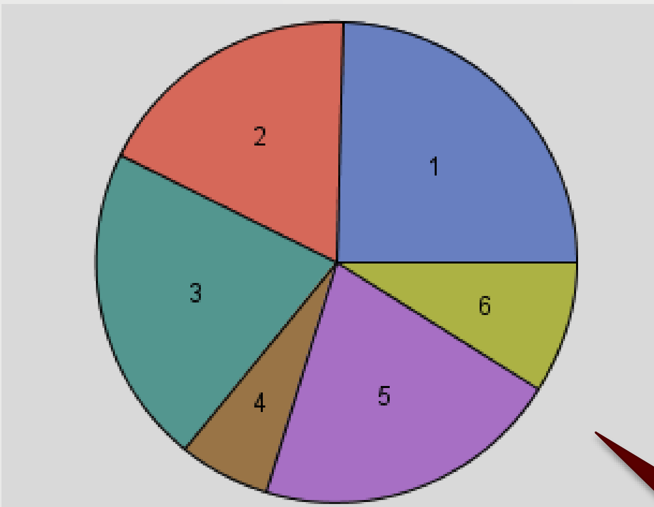
Default Properties

Common Property Changes:

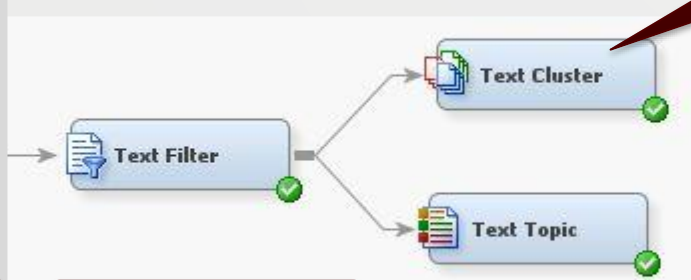
1. Maximum to Exact
2. Number of Clusters to less than 10
3. SVD Resolution to Medium

Cluster Analysis

✔ Results from using default properties



6 topics



Cluster Analysis

Topic's Word Description

Clusters

Cluster ID	Descriptive Terms	Frequency
1	encore wynn +floor beautiful +service +late +restaurant +credit +desk +wait +area +pool 'front desk' +hour +en...	408
2	excalibur +upgrade +clean tower +tower +location +value +price +casino +great +pay +money +kid +night sout...	309
3	bellagio +view +fountain oct beautiful +show +pool +restaurant +staff +hotel great +service +bathroom +buffet ...	357
4	+desk front +charge 'front desk' +credit +card +late +check +wait +know +line +hour +back +day +book ...	100
5	circus +kid +value +cheap strip +place +money +good +buffet +price +hotel +tower +end +walk +clean ...	347
6	bally north paris south tower +location +tower +upgrade +clean great +value +great +casino +walk +look ...	150

Cluster Analysis

✔ Example – Six Topic Clusters using SVD

Cluster Analysis Description

Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4
408	24%	0.586512	0.129415	0.039732	-0.03933
309	18%	0.617309	-0.20197	0.008146	0.053029
357	21%	0.628894	-0.02962	0.134059	-0.04731
100	6%	0.55446	-0.03027	-0.1421	-0.08149
347	21%	0.575499	-0.27352	-0.01744	-0.03493
150	9%	0.569563	-0.22299	0.0519	-0.23496

Number of Documents for Each of the 6 Topics

Total Number is 1,671 The number of Reviews

SVD Vectors (U)

Applied Analytics

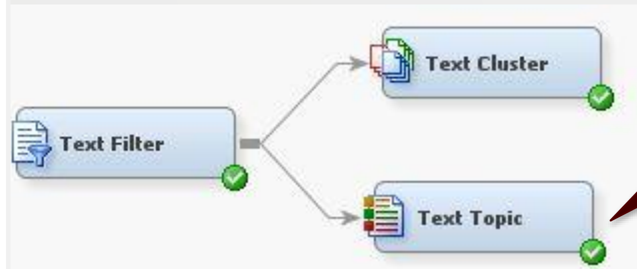
Text Topic Node

STAT 656 – Applied Analytics using SAS© Enterprise Miner™

Text Topic Node

✔ Text Topic Node Default Properties

General	
Node ID	TextTopic
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
User Topics	...
⊟ Term Topics	
Number of Single-term Topics	0
⊟ Learned Topics	
Number of Multi-term Topics	25
Correlated Topics	No
⊟ Results	
Topic Viewer	...
Status	
Create Time	7/20/16 9:02 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No



Text Topic Node

Default Properties

Interactive Topic Viewer

Common Property Changes:

1. Number of Multi-Term Topics
2. Correlated Topics
3. Use topic viewer to create Relevant Topics

Text Topic Node

✔ Results from using default properties

Topics

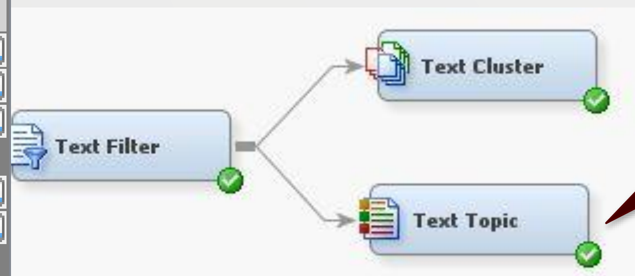
Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.060	0.021	+coffee,+suite,+maker,+buffet,+breakfast	356	167
Multiple	2	0.087	0.020	wynn,encore,encore,+property,xs	286	201
Multiple	3	0.089	0.021	+desk,front,front desk,+line,+manager	338	233
Multiple	4	0.092	0.020	bally,paris,tower,north,south	248	180
Multiple	5	0.079	0.020	excalibur,luxor,york,mandalay,mgm	296	187
Multiple	6	0.100	0.020	bellagio,+fountain,+show,beautiful,+view	292	247
Multiple	7	0.080	0.021	+suite,+tv,+bathroom,+shower,+tub	381	224
Multiple	8	0.072	0.021	+play,+drink,+slot,+poker,+eat	391	227
Multiple	9	0.113	0.020	+value,+great,great value,excellent,good value	271	239
Multiple	10	0.071	0.020	amp,+clean,+customer,+guest,+travel	208	140
Multiple	11	0.070	0.020	quot,+suite,awesome,+experience,+customer	253	303
Multiple	12	0.095	0.021	+review,+read,+love,+check,+check	351	271
Multiple	13	0.078	0.020	+smoke,+smoke room,+non-smoke,smoke,+smell	299	133
Multiple	14	0.065	0.021	+credit,+charge,+card,+charge,free	344	187
Multiple	15	0.089	0.020	circus,circus,west,+manor,first	257	168
Multiple	16	0.060	0.021	+bus,+ticket,+walk,+walk,+taxi	411	171
Multiple	17	0.073	0.020	+hot,+water,hot water,+spa,+shower	284	149
Multiple	18	0.057	0.021	+suite,las,bally,+staff,vegas	436	217
Multiple	19	0.079	0.020	circus,+kid,circus circus,vegas [illegible]amily	261	175
Multiple	20	0.078	0.021	+place,+pay,+cheap,+hotel,s[illegible]	372	261
Multiple	21	0.075	0.020	+tower,north,north tower,so[illegible] tower	306	136
Multiple	22	0.060	0.021	+sh[illegible]	418	197
Multiple	23	0.067	0.021	+pa[illegible]	375	232
Multiple	24	0.069	0.021	+ar[illegible]	388	243
Multiple	25	0.065	0.021	ball[illegible]	337	183

Default Number of Identified Topics (25)

Text Topic Node

✔ Text Topic Node Modified Properties

Property	Value
General	
Node ID	TextTopic
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
User Topics	...
Term Topics	
Number of Single-term Topics	0
Learned Topics	
Number of Multi-term Topics	7
Correlated Topics	No
Results	
Topic Viewer	...
Status	
Create Time	7/20/16 9:02 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No



Text Topic Node

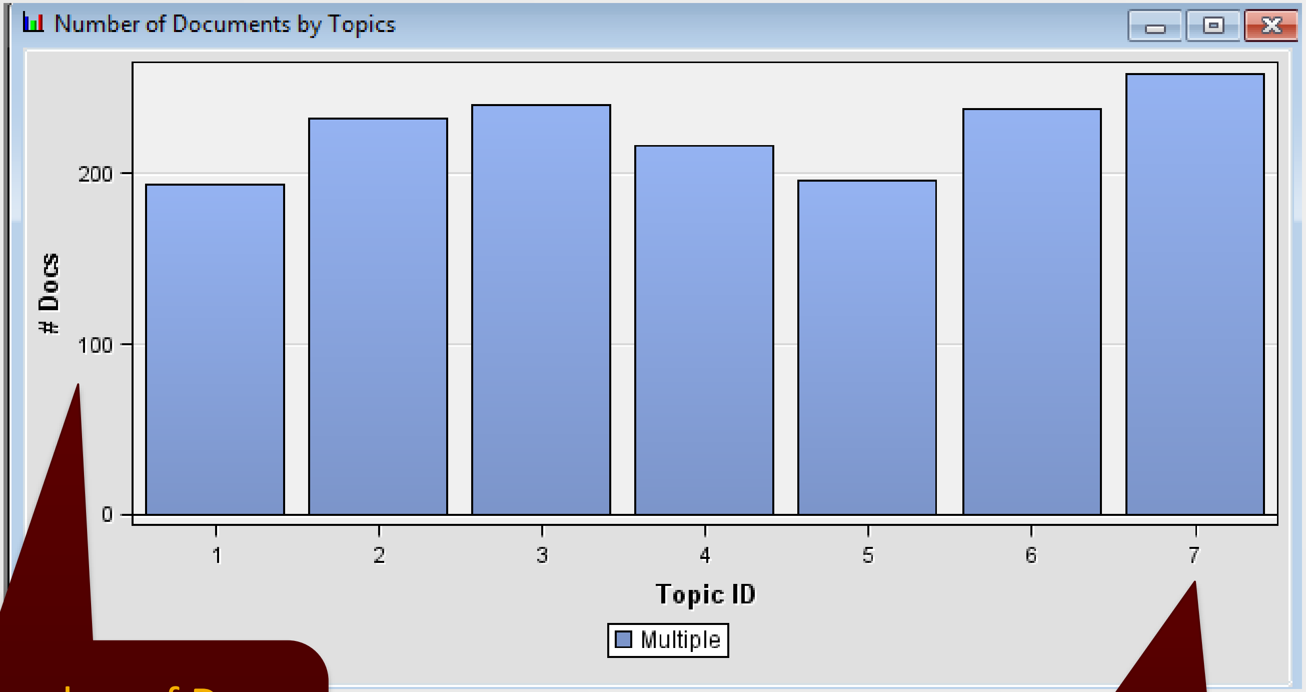
Changed Number of Topics from 25 to 7

Cluster Analysis Identified 6 topics

In general set the initial number of multi-term topics to approximately the number of topics found with Text Cluster.

Text Topic Node

✔ Results from using default properties



Number of Docs Associated with Each Topic

Initial Number of Identified Topics (7)

Text Topic Node

✔ Results from Restricting Number of Topics to 7

Topic ID	Topic	Document Cutoff	Term Cutoff	Number of Terms	# Docs
1	bally,paris,north,+tower,+location	0.136	0.020	257	193
2	wynn,encore,encore,+pool,+spa	0.110	0.020	287	232
3	quot,+desk,front,front desk,+manager	0.111	0.021	358	240
4	circus,circus,+kid,circus circus,strip	0.120	0.020	260	216
5	excalibur,york,luxor,mgm,mandalay	0.104	0.020	303	196
6	bellagio,+fountain,+view,beautiful,+amaze	0.118	0.020	313	238
7	+bed,+bathroom,+tv,+shower,+area	0.079	0.021	39[illegible]	259

Initial Number of Identified Topics (7)

Total Number is 1,574 a Little Less than Number of Reviews

Number of Docs Associated with Each Topic

Text Topic Node

✔ Topic Viewer – Used to Shape Topics

Interactive Topic Viewer

File Edit

Topics

Recalculate

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
bally,paris,north,+tower,+location	Multiple	**0.02**	**0.136**	257	193
excalibur,york,luxor,mgm,mandalay	Multiple	**0.02**	**0.104**	303	196
circus,circus,+kid,circus circus,strip	Multiple	**0.02**	**0.12**	260	216
wynn,encore,encore,+pool,+spa	Multiple	**0.02**	**0.11**	287	232
bellagio,+fountain,+view,beautiful,+amaze	Multiple	**0.02**	**0.118**	313	238
quot,+desk,front,front desk,+manager	Multiple	**0.021**	**0.111**	358	240
+bed,+bathroom,+tv,+shower,+area	Multiple	**0.021**	**0.079**	392	259

Topic Viewer

Terms

Topic Weight	+	Term	Role	# Docs	Freq
0.305		quot	Noun	354	1162
0.155	+	desk	Noun	[illegible]	[illegible]
0.152		front	Adj	172	[illegible]
0.144		front desk	Noun Group	136	176
0.132	+	manager	Noun	65	88
0.119	+	customer	Noun	93	122
0.115	+	check	Verb	392	550
0.113	+	card	Noun	82	126
0.11	+	credit	Noun	66	111

Term Viewer

Document Viewer

Documents

Topic Weight	Filtered	Accessed	Created	Extension	Filtered Size	Language	Modified	Name	Omitted	Size	Text	Truncated	URI
0.334	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	3386.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	3386.0	Jun 2 2009 Welcome to	0.0	file://C:\EM_Projects\H
0.328	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	4338.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	4350.0	Jun 21 2009 The Fish	0.0	file://C:\EM_Projects\H
0.328	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	3157.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	3157.0	Aug 13 2009 Great	0.0	file://C:\EM_Projects\H
0.314	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	7138.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	7138.0	Jun 22 2009 Glitchy My	0.0	file://C:\EM_Projects\H
0.299	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	2402.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	2402.0	Nov 2 2009 Encore	0.0	file://C:\EM_Projects\H
0.299	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	2217.0	English	2015-04-29 20:05:20.0	bellagio_las_vegas164.t	0.0	2217.0	Oct 2 2009 Overpriced I	0.0	file://C:\EM_Projects\H
0.288	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	2460.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	2460.0	Feb 4 2009 Bally's	0.0	file://C:\EM_Projects\H
0.284	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	1655.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	1655.0	Jun 22 2009 TERRIBLE	0.0	file://C:\EM_Projects\H
0.283	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	3979.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	3979.0	Jun 15 2009 Catch 22	0.0	file://C:\EM_Projects\H

Text Topic Node

✔ Topic Viewer – Used to Shape Topics

Interactive Topic Viewer

File Edit

Topics

Recalculate

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
bally,paris,north,+tower,+location	Multiple	0.02	0.136	257	193
excalibur,york,luxor,mgm,mandalay	Multiple	0.02	0.104	303	196
circus,circus,+kid,circus circus,strip	Multiple	0.02	0.12	260	216
wynn,encore,encore,+pool,+spa	Multiple	0.02	0.11	287	232
bellagio,+fountain,+view,beautiful,+amaze	Multiple	0.02	0.118	313	238
quot,+desk,front,front desk,+manager	Multiple	0.021	0.111	358	240
+bed,+bathroom,+tv,+shower,+area	Multiple	0.021	0.079	392	259

Initial 7 Topics

Terms

Topic Weight	+	Term	Role	# Docs	Freq
0.305		quot	Noun	354	1162
0.155	+	desk	Noun	[illegible]	341
0.152		front	Adj	172	[illegible]
0.144		front desk	Noun Group	136	176
0.132	+	manager	Noun	65	88
0.119	+	customer	Noun	93	122
0.115	+	check	Verb	392	550
0.113	+	card	Noun	82	126
0.11	+	credit	Noun	66	111

Initial Term Weights

Individual Docs /Reviews

Documents

Topic Weight	Filtered	Accessed	Created	Extension	Filtered Size	Language	Modified	Name	Omitted	Size	Text	Truncated	URI
0.334	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	3386.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	3386.0	Jun 2 2009 Welcome to	0.0	file://C:\EM_Projects\H
0.328	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	4338.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	4350.0	Jun 21 2009 The Fish	0.0	file://C:\EM_Projects\H
0.328	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	3157.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	3157.0	Aug 13 2009 Great	0.0	file://C:\EM_Projects\H
0.314	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	7138.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	7138.0	Jun 22 2009 Glitchy My	0.0	file://C:\EM_Projects\H
0.299	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	2402.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	2402.0	Nov 2 2009 Encore	0.0	file://C:\EM_Projects\H
0.299	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	2217.0	English	2015-04-29 20:05:20.0	bellagio_las_vegas164.t	0.0	2217.0	Oct 2 2009 Overpriced I	0.0	file://C:\EM_Projects\H
0.288	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	2460.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	2460.0	Feb 4 2009 Bally's	0.0	file://C:\EM_Projects\H
0.284	C:\EM_Projects\HW8-Se	2016-07-15 23:02:32.0	2016-07-15 23:02:32.0	.txt	1655.0	English	2015-04-29 20:05:20.0	encore_at_wynn_las_ve	0.0	1655.0	Jun 22 2009 TERRIBLE	0.0	file://C:\EM_Projects\H
0.283	C:\EM_Projects\HW8-Se	2016-07-15 23:02:31.0	2016-07-15 23:02:31.0	.txt	3979.0	English	2015-04-29 20:05:20.0	bally_s_las_vegas_hotel	0.0	3979.0	Jun 15 2009 Catch 22	0.0	file://C:\EM_Projects\H

Text Topic Node

✔ Select a Topic to View Associated Documents

Individual Customer Reviews

Name	Omitted	Size	Text
bally_s_las_vegas_hotel_casino187.txt	0.0	3386.0	Jun 2 2009 Welcome to
encore_at_wynn_las_vegas236.txt	0.0	4350.0	Jun 21 2009 The Fish
encore_at_wynn_las_vegas151.txt	0.0	3157.0	Aug 13 2009 Great
encore_at_wynn_las_vegas234.txt	0.0	7138.0	Jun 22 2009 Glitchy My
encore_at_wynn_las_vegas30.txt	0.0	2402.0	Nov 2 2009 Encore
bellagio_las_vegas164.txt	0.0	2217.0	Oct 2 2009 Overpriced I
bally_s_las_vegas_hotel_casino306.txt	0.0	2460.0	Feb 4 2009 Bally's
encore_at_wynn_las_vegas232.txt	0.0	1655.0	Jun 22 2009 TERRIBLE

Top 3 Reviews Representing Topic 6

Topic 6 Documents Ordered by Relevance

Name	.	.	Text
bally_s_las_vegas_hotel_casino187.txt	0.0	3386.0	Jun 2 2009 Welcome to the worst of the strip I booked a flight+stay package thru Allegiant air 2 months before my trip, they made a mistake when booking the hotel though which I found out at 11:00 pm when I arrived to the hotel after waiting 35 minutes in line, they blamed Allegiant for the mistake without offering any solution. The manager called "LARRY" came and made us feel that it was our fault he said, "...we had NOTHING available..." even though the kid who was helping us first said that if we wanted to stay we had to pay $219.00 per night. Of course that's the price for a "high end" suite, the helpful manager wouldn't want us to stay in a expensive suite but please, don't say there is nothing available. One hour later and without receiving any help at all (The manager didn't call another hotel) we grabbed our bags and started walking towards the Monte Carlo.Next day
encore_at_wynn_las_vegas236.txt	0.0	4350.0	Jun 21 2009 The Fish Stinks from the Head My wife and I just returned from a three night stay at Encore (6/17 6/20). The stay started out well. Check-in went smoothly and we upgraded to a slightly larger room on the 53rd floor that was nicely appointed and included a panoramic view of Las Vegas through floor to ceiling windows. Being a new hotel, the room had some nice amenities, such as bedside controls for the blinds and curtains. You can also request privacy or maid service from the bedside controls.As other reviewers have stated, the pool experience is subpar, even for a lesser hotel. The waitress service in the euro deck area was sometimes non existent and we had to go to the bar three times ourselves to get drinks and food. They were out of several food items and even some of the supplies need to make all of the standard drinks on the menu. When you do get the food, it is disappointing to see that it is
encore_at_wynn_las_vegas151.txt	0.0	3157.0	Aug 13 2009 Great hotel but not great service My husband and I invited some family to stay at the Encore for one week. In the past we've been at the Wynn many times, at least once/year since it opened.Since they were our guests, all rooms were recorded under his name.The last day we wanted to treat the women of the group with a massage. We booked the spa treatments, all five of us, under my husband's last name, giving different first names, because it was simpler. The lady at the reception - Yulia - informed us that my husband was the only one who could charge the treatments to the room, but that we could have paid with credit cards. No problem.At the time of the treatment, coming directly from the pool where we spent the day, we were informed we needed IDs to be able to pay with the credit card. The same credit card had been accepted in the restaurants during the week, to pay for dinners much more

Text Topic Node

✔ Label Topic With Meaningful Description

Topic Descriptions

Document Cutoff

Number of Documents

Topics

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
bally,paris,north,+tower,+location	Multiple	**0.02**	**0.136**	257	193
excalibur,york,luxor,mgm,mandalay	Multiple	**0.02**	**0.104**	303	196
circus,circus,+kid,circus circus,strip	Multiple	**0.02**	**0.12**	260	216
wynn,encore,encore,+pool,+spa	Multiple	**0.02**	**0.11**	287	232
bellagio,+fountain,+view,beautiful,+amaze	Multiple	**0.02**	**0.118**	313	238
Desk and Mangement Problems	User	**0.021**	**0.111**	358	240
+bed,+bathroom,+tv,+shower,+area	Multiple	**0.021**	**0.079**	392	259

User Described Topic

Category Automatically Changes to User

Text Topic Node

✔ Labeled 4 USER Topics

Topic Descriptions

The total number of Docs from all topics should be close to the total number of Docs in the Data (N=1,671)

Topics

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
bally,paris,north,+tower,+location	Multiple	**0.02**	**0.136**	257	193
excalibur,york,luxor,mgm,mandalay	Multiple	**0.02**	**0.104**	303	196
Circus Circus	User	**0.02**	**0.12**	260	216
wynn,encore,encore,+pool,+spa	Multiple	**0.02**	**0.11**	287	232
Bellagio	User	**0.02**	**0.118**	313	238
Desk and Mangement Problems	User	**0.021**	**0.111**	358	240
Accomodations	User	**0.021**	**0.079**	392	259

Total = 1,574

User Described Topic

4 USER Topics

The SVD score for a document must be above the Cutoff for it to be considered in that topic.

Text Topic Node

✔ Save USER Topics and Request More

Request only 2-3 New Multiple Topics

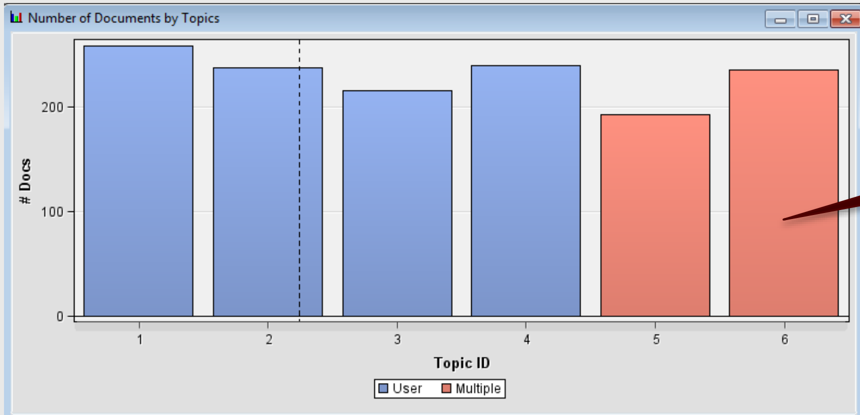
Near the End, Allow Correlated Topics

Return to the Topic Viewer to See New Topics

Property	Value
General	
Node ID	TextTopic
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
User Topics	...
Term Topics	
Number of Single-term Topics	0
Learned Topics	
Number of Multi-term Topics	2
Correlated Topics	No
Results	
Topic Viewer	...
Status	
Create Time	7/20/16 9:02 PM
Run ID	3dd9f150-948b-45cd-8773-04c5
Last Error	
Last Status	Complete
Last Run Time	7/20/16 9:04 PM
Run Duration	0 Hr. 0 Min. 3.20 Sec.
Grid Host	
User-Added Node	No

Text Topic Node

✔ Review New Topics Using Topic Viewer



Two New Topics from SVD

Topic ID	Topic	# Docs
1	Accomodations	259
2	Bellliagio	238
3	Circus Circus	216
4	Front Desk Issues	240
5	excalibur,york,luxor,mgm,+pool	193
6	wynn,encore,encore,+pool,+area	235

Previously Defined USER Topics

Review and Shape New Multi-Term Topics

In this case, new topics are similar to previous topics. In general they can change.

Text Topic Node

✔ All User Topics

Topic	Category	# Docs	Term Cutoff	Document Cutoff	Number of Terms
Excalibur	User	193	**0.02**	**0.106**	314
Circus Circus	User	216	**0.02**	**0.12**	260
Encore	User	235	**0.02**	**0.11**	290
Bellliagio	User	238	**0.02**	**0.118**	313
Front Desk Issues	User	240	**0.021**	**0.111**	358
Accomodations	User	259	**0.021**	**0.079**	392

All USER Topics

Check Total Number of Docs to See If This is Close to Number in Data (N=1,671)

Text Topic Node

✔ Cluster Topics Vs. Text Topic Node

Cluster ID	Descriptive Terms
1	circus +kid +cheap +place +end +price +tower +money strip +buffet +walk las +good +look +hotel ...
2	bally +tower +location north paris south +clean +upgrade great +great strip +walk +stay +time +casino ...
3	+coffee south +suite +food +end +floor +look +upgrade bally paris +service north +bathroom +book +day ...
4	bellagio +fountain beautiful +view excellent +show +staff +hotel +restaurant +buffet great +service +great +pool ...
5	encore wynn beautiful +suite +service +restaurant +floor +pool +area las +view excellent +staff +coffee first ...
6	excalibur +kid quot +money +pay +day +bed +buffet +want +casino first +check +upgrade +cheap +price ...

Cluster Analysis Topics

Text Topic Node Solution

Topic	Category	# Docs	Term Cutoff	Document Cutoff	Number of Terms
Excalibur	User	193	**0.02**	**0.106**	314
Circus Circus	User	216	**0.02**	**0.12**	260
Encore	User	235	**0.02**	**0.11**	290
Bellliagio	User	238	**0.02**	**0.118**	313
Front Desk Issues	User	240	**0.021**	**0.111**	358
Accomodations	User	259	**0.021**	**0.079**	392

Week 9

Next Week

Next

Topic Analysis and Document Classification

Sentiment Analysis (Week 10)

Final Exam (Week 11)