**SAS ENTERPRISE MINER AND PYTHON-APPENDICES**

[**1. Topic Analysis**](#_ukhzolk1kqq0) **2**

[**2. Parsing and Filtering Sentiment Words**](#_61p08stnp4xe) **13**

[**3. Calculating Text Sentiment**](#_hgf5e4tge8n4) **16**

[**4. 10 Fold Cross Validation (Topic Node)**](#_ym9x3x9gnx4i) **19**

[**5. 10 Fold Cross Validation (Cluster Node)**](#_k0cq3vr66d4p) **25**

[**6. Python**](#_r9uxltok2b43) **36**

# **1. Topic Analysis**

A computer screen shot of a diagram

Description automatically generated

File Import

A screenshot of a computer

Description automatically generated

*Note: Reject “State”, put “Crash” as Target*

A screenshot of a computer

Description automatically generated

*Note: Set “Fetch Size” to Max. “Mileage” is missing*

Stat Explorer:

A close-up of words

Description automatically generated

Since the Probability of Crash is more than 10%, this is not a Rare Event Problem

Replacement:

A screenshot of a computer

Description automatically generated

A screenshot of a computer error

Description automatically generated

A screenshot of a computer

Description automatically generated

A close-up of several words

Description automatically generated

Impute Node:

A screenshot of a computer

Description automatically generated

A group of black text

Description automatically generated

Text Parsing

A screenshot of a computer

Description automatically generated

Text Filter:

A screenshot of a computer

Description automatically generated

*Filter: Raw Count-IDF. Terms must appear in at least 4 documents to be counted.*

*A table of numbers and letters

Description automatically generated with medium confidence*

Text Cluster:

A screenshot of a computer

Description automatically generated

*7 Clusters - Low:*

*A pie chart with numbers

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

(1: Beam - Light, 2: Airbag, 4: Transmission + Gear, 5: srs lights + seat, 6: Brake +Wheel, 7: Tire)

*7 Clusters Medium:*

*A pie chart with numbers

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

(1: Tire, 3: Mileage, 4: Transmission + Engine, 5: Beam + Light, 6: Airbag, 6: srs light+airbag)

*7 Clusters - High*

*A circular diagram with numbers

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

(1: Tire, 3/5: transmission + engine, 4: mileage, 7: airbag + srs light)

**=> Take out additional stop words.**

I do a second parsing to remove further stop words that seem irrelevant to the topic assignment: “problem”,”contact”, “vehicle”,”dealer”,”car”, “consumer”,”issue”, “failure”

Text Cluster:

*7 Clusters - Low*

*A pie chart with numbers

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

*(1: Light, 2: Door+Wheel, 3: Airbag, 4: Transmission, 5: Tire, 6: Mileage, 7: Beam+Light)*

*7 Clusters - Medium*

*A pie chart with numbers

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

(1: Beam+Light, 2: Door, 3: Mileage, 4: Transmission, 5: Airbag, 6: srs ligh, 7 brake+ignition)

*7 Clusters- High*

*A pie chart with numbers and a few different colored circles

Description automatically generated with medium confidence*

*A screenshot of a computer

Description automatically generated*

Text Topic

A screenshot of a computer

Description automatically generated

First time: Set Correlated Topics to No. After that, set it to Yes. In the final run, set “Number of Multi-term Topics” to 0

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **2. Parsing and Filtering Sentiment Words**

Text Sentiment Parsing

A screenshot of a computer

Description automatically generated

Text Sentiment Filter

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# 

# 

# 

# 

# 

# 

# **3. Calculating Text Sentiment**

A screenshot of a computer

Description automatically generated

SAS CODE: Text Filter TMOUT

A screenshot of a computer code

Description automatically generated

SAS CODE: Text Filter Graph Table

A screenshot of a computer code

Description automatically generated

Merge Node:

A screenshot of a computer

Description automatically generated

A screenshot of a computer error

Description automatically generated

Save Data:

A screenshot of a computer

Description automatically generated

SAS: Calculate and Save Sentiment

proc sort data=mylib.merge\_train;

by term; run;

data mylib.TermDocMatrix;

merge mylib.merge\_train mylib.afinn\_sentimentscore;

by Term;

keep Term \_Document\_ Termnumber Count Score;

if Score ne . and Term ne MISSING and \_Document\_ ne . then output;

run;

proc sort data=mylib.TermDocMatrix; by \_Document\_ Term; run;

Data mylib.sentiment;

retain Docscore n; keep \_Document\_ n Docscore stars;

set mylib.TermDocMatrix; by \_Document\_ Term;

if first.\_Document\_ then do;

Docscore = count\*score;

n = count;

end;

else do;

docscore= docscore + count\*score;

n= n + 1;

end;

if last.\_Document\_ then do;

if n>0 then docscore=docscore/n;

else docscore=0;

if Docscore ne MISSING then stars = 3+ (4/6)\*Docscore; output;

end;

run;

data mylib.sentiment;

retain Doc 0 nsave docscoreSave starsSave DocSave;

KEEP \_Document\_ n DocScore Stars;

set mylib.Sentiment;

doc= doc+1;

if doc lt \_Document\_ then do;

nSave=n; DocScoreSave= DocScore; StarsSave=Stars;

DocSave=\_Document\_;

do while (doc LT DocSave);

n=0; DocScore=0; Stars=3; \_Document\_=Doc;

output; doc=doc+1;

end;

n=nSave; DocScore=DocScoreSave; Stars=StarsSave; \_Document\_=DocSave;

end;

if DocScore eq . then do; n=0; DocScore=0; Stars=3;

end; output;

run;

proc means data=mylib.sentiment;

var Docscore;

run;

A close-up of a document

Description automatically generated

# **4. 10 Fold Cross Validation (Topic Node)**

A screenshot of a computer

Description automatically generated

Input: Topic\_Train

A screenshot of a computer

Description automatically generated

Merge: Topic and Sentiment

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

SAS CODE: 10 FOLDS (outside loop)

A screenshot of a computer

Description automatically generated

data mylib.selection;

call streaminit(12345);

set &EM\_IMPORT\_DATA; urand=rand('uniform');

proc sort data=mylib.selection; by urand;

data &EM\_EXPORT\_TRAIN;

drop fold\_size urand;

set mylib.selection NOBS=nobs\_;

fold\_size=round(nobs\_/10);

if \_N\_ <=fold\_size then fold ='A'; if \_N\_>fold\_size and \_N\_<=fold\_size\*2 then fold = 'B';

if \_N\_> 2\*fold\_size and \_N\_<=fold\_size\*3 then fold = 'C';

if \_N\_> 3\*fold\_size and \_N\_<=fold\_size\*4 then fold = 'D';

if \_N\_> 4\*fold\_size and \_N\_<=fold\_size\*5 then fold = 'E';

if \_N\_> 5\*fold\_size and \_N\_<=fold\_size\*6 then fold = 'F';

if \_N\_> 6\*fold\_size and \_N\_<=fold\_size\*7 then fold = 'G';

if \_N\_> 7\*fold\_size and \_N\_<=fold\_size\*8 then fold = 'H';

if \_N\_> 8\*fold\_size and \_N\_<=fold\_size\*9 then fold = 'I';

else if \_N\_>9\*fold\_size then fold='J';

proc freq data=&EM\_EXPORT\_TRAIN;

by fold; tables crash;

run;

Metadata:

A screenshot of a computer

Description automatically generated

Save Node:

A screenshot of a computer

Description automatically generated

Start Group:

A screenshot of a computer

Description automatically generated

SAS CODE In Loop

A screenshot of a computer

Description automatically generated

data mylib.topic\_temp1;

retain c1-c10 0; keep c1-c10;

set &EM\_IMPORT\_DATA end=eof;

if fold= 'A' then c1= c1+1; if fold= 'B' then c2= c2+1;

if fold= 'C' then c3= c3+1; if fold= 'D' then c4= c4+1;

if fold= 'E' then c5= c5+1; if fold= 'F' then c6= c6+1;

if fold= 'G' then c7= c7+1; if fold= 'H' then c8= c8+1;

if fold= 'I' then c9= c9+1; if fold= 'J' then c10= c10+1;

if eof then output;

data &em\_export\_validate;

drop c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 rfold;

retain rfold '0';

set mylib.TopicAll\_Train; if rfold='0' then do;

set mylib.topic\_temp1;

if c1=0 then rfold ='A'; if c2=0 then rfold ='B';

if c3=0 then rfold ='C'; if c4=0 then rfold ='D';

if c5=0 then rfold ='E'; if c6=0 then rfold ='F';

if c7=0 then rfold ='G'; if c8=0 then rfold ='H';

if c9=0 then rfold ='I'; if c10=0 then rfold ='J';

end; if fold=rfold then output; run;

*(Note: the data name for each loop is different. The names are topic\_temp1 to topic\_temp 7)*

In-loop Decision Tree (Maximum Depths: 5,6,8,10,12,15, and 17)

data mylib.topic\_temp1;

retain c1-c10 0; keep c1-c10;

set &EM\_IMPORT\_DATA end=eof;

if fold= 'A' then c1= c1+1; if fold= 'B' then c2= c2+1;

if fold= 'C' then c3= c3+1; if fold= 'D' then c4= c4+1;

if fold= 'E' then c5= c5+1; if fold= 'F' then c6= c6+1;

if fold= 'G' then c7= c7+1; if fold= 'H' then c8= c8+1;

if fold= 'I' then c9= c9+1; if fold= 'J' then c10= c10+1;

if eof then output;

data &em\_export\_validate;

drop c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 rfold;

retain rfold '0';

set mylib.TopicAll\_Train; if rfold='0' then do;

set mylib.topic\_temp1;

if c1=0 then rfold ='A'; if c2=0 then rfold ='B';

if c3=0 then rfold ='C'; if c4=0 then rfold ='D';

if c5=0 then rfold ='E'; if c6=0 then rfold ='F';

if c7=0 then rfold ='G'; if c8=0 then rfold ='H';

if c9=0 then rfold ='I'; if c10=0 then rfold ='J';

end; if fold=rfold then output; run;

Confusion Matrix

A table with numbers and a yellow box

Description automatically generated

We can see here that increasing the tree depth beyond 15 does not improve the matrix. Therefore, I stopped at depth 15.

# **5. 10 Fold Cross Validation (Cluster Node)**

A screenshot of a computer

Description automatically generated

Merge Node:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

SAS CODE (Outside Loop)

A screenshot of a computer

Description automatically generated

data mylib.selection2;

call streaminit(12345);

set &EM\_IMPORT\_DATA; urand=rand('uniform');

proc sort data=mylib.selection; by urand;

data &EM\_EXPORT\_TRAIN;

drop fold\_size urand;

set mylib.selection2 NOBS=nobs\_;

fold\_size=round(nobs\_/10);

if \_N\_ <=fold\_size then fold ='A'; if \_N\_>fold\_size and \_N\_<=fold\_size\*2 then fold = 'B';

if \_N\_> 2\*fold\_size and \_N\_<=fold\_size\*3 then fold = 'C';

if \_N\_> 3\*fold\_size and \_N\_<=fold\_size\*4 then fold = 'D';

if \_N\_> 4\*fold\_size and \_N\_<=fold\_size\*5 then fold = 'E';

if \_N\_> 5\*fold\_size and \_N\_<=fold\_size\*6 then fold = 'F';

if \_N\_> 6\*fold\_size and \_N\_<=fold\_size\*7 then fold = 'G';

if \_N\_> 7\*fold\_size and \_N\_<=fold\_size\*8 then fold = 'H';

if \_N\_> 8\*fold\_size and \_N\_<=fold\_size\*9 then fold = 'I';

else if \_N\_>9\*fold\_size then fold='J';

proc freq data=&EM\_EXPORT\_TRAIN;

by fold; tables crash;

run;

Metadata Node:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Save Node:

A screenshot of a computer

Description automatically generated

Start Node:

A screenshot of a computer

Description automatically generated

In-loop SAS Code:

A screenshot of a computer

Description automatically generated

data mylib.cluster\_temp1;

retain c1-c10 0; keep c1-c10;

set &EM\_IMPORT\_DATA end=eof;

if fold= 'A' then c1= c1+1; if fold= 'B' then c2= c2+1;

if fold= 'C' then c3= c3+1; if fold= 'D' then c4= c4+1;

if fold= 'E' then c5= c5+1; if fold= 'F' then c6= c6+1;

if fold= 'G' then c7= c7+1; if fold= 'H' then c8= c8+1;

if fold= 'I' then c9= c9+1; if fold= 'J' then c10= c10+1;

if eof then output;

data &em\_export\_validate;

drop c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 rfold;

retain rfold '0';

set mylib.ClusterAll\_Train; if rfold='0' then do;

set mylib.cluster\_temp1;

if c1=0 then rfold ='A'; if c2=0 then rfold ='B';

if c3=0 then rfold ='C'; if c4=0 then rfold ='D';

if c5=0 then rfold ='E'; if c6=0 then rfold ='F';

if c7=0 then rfold ='G'; if c8=0 then rfold ='H';

if c9=0 then rfold ='I'; if c10=0 then rfold ='J';

end; if fold=rfold then output; run;

*(the data name varies from cluster\_temp1 to cluster\_temp6)*

Decision Tree:

A screenshot of a computer

Description automatically generated

*(The tree depths ranges from 6,8,10,12,14, and 16)*

Confusion Matrix

A table with numbers and a yellow box

Description automatically generated

We can see here that increasing the tree depth beyond 14 does not improve the matrix. Therefore, I stopped at depth 14.

A screenshot of a computer flowchart

Description automatically generated

MetaData (Cluster)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Metadata (Topic)

A screenshot of a computer

Description automatically generated

Non-HP Partition

A screenshot of a table

Description automatically generated

Non-HP Decision Trees

A screenshot of a computer

Description automatically generated

*(Note: Maximum Depth is 15 for Topic and 14 for Cluster)*

HP Partition Nodes:

A screenshot of a computer

Description automatically generated

HP Trees:  
A screenshot of a computer

Description automatically generated

*(Note: Maximum Depth is 15 for Topic and 14 for Cluster)*

Confusion Matrix

When Using Gini for all Trees

A screenshot of a graph

Description automatically generated

When using Entropy, the result for Topic Non Hp is better

A screenshot of a graph

Description automatically generated

When using ProbChiSq for Non-Hp Trees and Chi-Square for Hp Trees, the results got worse

A screenshot of a graph

Description automatically generated

=> **Best Trees: Topic Non HP tree** using Entropy

A diagram of a computer

Description automatically generated with medium confidence

A computer screen shot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

We can see that the topic of Airbags has the highest importance in predicting the crash probability, followed by mph.

# **6. Python**

A screenshot of a computer code

Description automatically generated

Data obtained after the topic analysis and sentiment analysis

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated