TODAY’S AGENDA

INTRODUCTION TO SAS® TEXT MINER™

- Define data mining
- Overview of SAS® Enterprise Miner™
- Describe text analytics and define text data mining
- Text Mining Process
- SAS® Text Miner™
- Illustrate text mining by example

- Q&A
A QUICK INTRODUCTION TO DATA MINING
Turning increasing amounts of raw data into useful information
DATA MINING IS:

- Discovering patterns, trends and relationships represented in data
- Developing models to understand and describe characteristics and activity based on these patterns
- Using insights to help evaluate future options and take fact-based decisions
- Deploying scores and results for timely, appropriate action that affects the bottom line
SAS® ENTERPRISE MINER™
MODEL DEVELOPMENT PROCESS

Sample
- Input Data
- File Import
- Sample
- Data Partition
- Merge
- Filter
- Append
- Time Series

Explore
- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Path Analysis

Modify
- Transform Variables
- DMDB
- SOM/Kohonen
- Graph Explore
- Interactive Binning
- Drop
- Principal Components

Model
- Decision Tree
- Neural Network
- AutoNeural
- SVM
- Partial Least Squares
- Regression
- Rule Induction
- Model Import

Assess
- Model Comparison
- Score
- Segment Profile
- Decisions
- Cutoff

Text
- Text Import
- Text Parsing
- Text Filter
- Text Topic
- Text Cluster
- Text Rule Builder
A GLIMPSE OF SAS® ENTERPRISE MINER™

SAMPLE  EXPLORE  MODIFY  MODEL  ASSESS and SCORE

S ample  E xplore  M odify  M odel  A ssess
Tools are organized to support the SEMMA process:
- Sample
- Explore
- Modify
- Model
- Assess
Tool Bar Shortcut Buttons

Project Panel

Properties Panel

Help Panel

Diagram Workspace

Diagram Navigation Toolbar
WHY MINE TEXT?

IS VALUABLE INFORMATION “LOCKED AWAY” IN UNSTRUCTURED DATA?
UNSTRUCTURED AND SEMI-STRUCTURED DATA

- Unstructured data: 70%
- Structured data: 25%
- Semistructured data: 5%
Structured Data

- Age Group = 60+
- Satisfaction = Not Very
- Rewards Customer = No
- Total Hold Time = 8

Unstructured Data

- they called me so I returned their call because it was cut off in the middle of the conversation. Every time they call me, they're cut off.
WHY MINE TEXT?

WHAT CAN BE LEARNED FROM UNSTRUCTURED DATA?

• Are any of these documents related to one another based on their contents and the characteristics of their contents?
• What are the key topics, themes or concepts being discussed?
• Are there emerging issues?
• Do the documents contain potentially valuable information that could improve predictive models?
The process of **discovering and extracting** meaningful patterns and relationships from text collections.
1. Pattern Discovery (Unsupervised Learning)
2. Prediction (Supervised Learning)

These are the same general goals of data mining.
TEXT MINING

THE PROCESS

- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis
When SAS® Text Miner is licensed, an additional tab, “Text Mining”, appears in the workspace, containing tools to process and analyze unstructured data.
TEXT MINING
THE PROCESS

Text Preprocessing
Text Parsing
Transformation (Dimension Reduction)
Document Analysis
The expected SAS data set for text mining should have the following characteristics:

- One row per document
- A document ID (suggested)
- A “text” column

The “text” column can be either:

- The actual full text of the document, up to 32,000 characters
- A pointer to a text file (*.txt, *.html) located on the file system

The SAS data set can also have structured data and a target variable (dependent variable, response variable)
TEXT IMPORT NODE

- Enables you to create data sets dynamically from files contained in a directory or from the Web.
- Takes an import directory containing text files in potentially proprietary formats such as MS Word and PDF files as input.
- Extracts the text from the files, places a copy of the text in a plain text file, and a snippet (or possibly even all) of the text in a SAS data set.
- If a URL is specified, the node will crawl Web sites and retrieve files from the Web.
- The output of a **Text Import** node is a data set that can be imported into the **Text Parsing** node.
<table>
<thead>
<tr>
<th>TOT_AST_HOLD_DUR</th>
<th>TOT_AG'TALK_DUR</th>
<th>CALL_REASON</th>
<th>CSAT_BRAND_RELATIONSHIP</th>
<th>ASAT_RESPONSE</th>
<th>CSAT_OVERALL</th>
<th>v_close...</th>
<th>v_call_reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>969.0</td>
<td>Stop Payment on an account</td>
<td>Very</td>
<td>Completely satisfied</td>
<td>Somewhat</td>
<td>i called them</td>
<td>just charges that were not supposed to be there was a problem with a charge</td>
</tr>
<tr>
<td>5.0</td>
<td>546.0</td>
<td>Statement Questions/billing issues</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>i was reviewing my statement made a charge on my credit card program trying to get a hold of the fraud department</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>1162.0</td>
<td>Dispute the Validity of a Fee</td>
<td>Somewhat</td>
<td>Very</td>
<td>Not very</td>
<td>i'm satisfied</td>
<td>there was a problem with a charge</td>
</tr>
<tr>
<td>13.0</td>
<td>263.0</td>
<td>Dispute A Merchant Charge</td>
<td>Somewhat</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>i'm still called trying to get a hold of a rep</td>
<td></td>
</tr>
<tr>
<td>75.0</td>
<td>573.0</td>
<td>Questions about Account Security</td>
<td>Somewhat</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>keep trying</td>
<td>trying to get a hold of the fraud department</td>
</tr>
<tr>
<td>51.0</td>
<td>686.0</td>
<td>Request Hardship</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>because i... to receive financial assistance</td>
<td></td>
</tr>
<tr>
<td>106.0</td>
<td>415.0</td>
<td>Advanced Payment/Set Up Pay...</td>
<td>Somewhat</td>
<td>Not very</td>
<td>Not very</td>
<td>i was called</td>
<td>had received several messages from...</td>
</tr>
<tr>
<td>66.0</td>
<td>810.0</td>
<td>Notify Of Late Payment</td>
<td>Somewhat</td>
<td>Very</td>
<td>Somewhat</td>
<td>i just called</td>
<td>because i told them i could</td>
</tr>
<tr>
<td>163.0</td>
<td>693.0</td>
<td>Make a Payment</td>
<td>Not very</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>they were trying to pay my monthly statements with...</td>
<td></td>
</tr>
<tr>
<td>69.0</td>
<td>1537.0</td>
<td>Request Hardship</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>Not very</td>
<td>i was acc... the reason was to let them know what...</td>
<td></td>
</tr>
<tr>
<td>154.0</td>
<td>1132.0</td>
<td>Haven't Received Statement</td>
<td>Somewhat</td>
<td>Completely satisfied</td>
<td>Somewhat</td>
<td>i didn't g... they changed my account number and...</td>
<td></td>
</tr>
<tr>
<td>101.0</td>
<td>880.0</td>
<td>Change or Inquiry regarding PA...</td>
<td>Somewhat</td>
<td>Somewhat</td>
<td>Not very</td>
<td>essential... kind of a lengthy reason, bottom line</td>
<td></td>
</tr>
<tr>
<td>76.0</td>
<td>454.0</td>
<td>Make a Payment</td>
<td>Somewhat</td>
<td>Not very</td>
<td>Somewhat</td>
<td>i didn't m... to pay off my credit card...</td>
<td></td>
</tr>
<tr>
<td>274.0</td>
<td>941.0</td>
<td>Dispute the Validity of a Fee</td>
<td>Not at all satisfied</td>
<td>Completely satisfied</td>
<td>Not very</td>
<td>just the... about the $25 fee that kept popping up...</td>
<td></td>
</tr>
<tr>
<td>35.0</td>
<td>597.0</td>
<td>Advanced Payment/Set Up Pay...</td>
<td>Very</td>
<td>Completely satisfied</td>
<td>Somewhat</td>
<td>there were... to get some kind of payment arrangement...</td>
<td></td>
</tr>
<tr>
<td>69.0</td>
<td>529.0</td>
<td>Request Hardship</td>
<td>Not very</td>
<td>Not at all satisfied</td>
<td>Not very</td>
<td>my wife... my wife and i are both unemployed, a...</td>
<td></td>
</tr>
<tr>
<td>50.0</td>
<td>376.0</td>
<td>Change or Inquiry regarding PA...</td>
<td>Somewhat</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>i was trying... just a lost of job and things were getting...</td>
<td></td>
</tr>
<tr>
<td>52.0</td>
<td>941.0</td>
<td>Advanced Payment/Set Up Pay...</td>
<td>Not very</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>i felt... to communicate about why i was still... to get caught up on my payments...</td>
<td></td>
</tr>
<tr>
<td>53.0</td>
<td>504.0</td>
<td>Make a Payment</td>
<td>Somewhat</td>
<td>Somewhat</td>
<td>Somewhat</td>
<td>the rep... they had called and i wanted to try to...</td>
<td></td>
</tr>
<tr>
<td>30.0</td>
<td>1407.0</td>
<td>Advanced Payment/Set Up Pay...</td>
<td>Somewhat</td>
<td>Completely satisfied</td>
<td>Somewhat</td>
<td>some thing... change of address...</td>
<td></td>
</tr>
<tr>
<td>53.0</td>
<td>660.0</td>
<td>Change Name/Address on Account</td>
<td>Not very</td>
<td>Not at all satisfied</td>
<td>Not at all satisfied</td>
<td>i guess it...</td>
<td></td>
</tr>
<tr>
<td>53.0</td>
<td>1032.0</td>
<td>Request Hardship</td>
<td>Somewhat</td>
<td>Very</td>
<td>Somewhat</td>
<td>i don't... the first time i called, someone from...</td>
<td></td>
</tr>
</tbody>
</table>
TEXT MINING PROCESS

- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis
Text parsing decomposes textual data and generates a quantitative representation suitable for data mining purposes.

It transforms this:

```
v_call_reason

just charges that were not supposed to be on the account.

i called trying to get a hold of one representative that's been very nice through the whole ordeal, that's been trying to help me and she gave me her employee id number and told me to contact her. that way, she could give it more a one-on-one instead of speaking to a hundred different people and they wouldn't put me through to her. they said it was impossible for them to put me through to her when she had said it wasn't impossible so i got nowhere with that conversation, on saturday.

there was a problem with a charge.

i made a charge on my credit card probably now close to a month ago to an auto mechanic shop. the mechanic shop did terrible work, so i wanted to dispute my charge, but i'm not being able to process my dispute.

trying to get a hold of the fraud department. contact who has my case. she's never there, never calls back.

to receive financial assistance.

i had received several messages from orion. they were trying to reach me for a payment for this month.

i called them because i told them i couldn't make a payment until the 6th of january.

i tried to pay my monthly statements via my smart phone, but i was told my customer service online technical support department that their online payment system is not compatible with smart phones.

the reason was to let them know what was going on in our life, and that we would not be able to pay this bill. i did make the last bill, i made the minimum payment. all i got out of orion. "if you pay next month's minimum payment, we'll give you $39, too." i says, "are you going to do that every month?" "oh no, we'll have to charge you a $25 late fee," to me, "you can save $25..." i says, "i call a little bit too!"
Text parsing decomposes textual data and generates a quantitative representation suitable for data mining purposes.

... into this:

<table>
<thead>
<tr>
<th>Terms</th>
<th>FREQ</th>
<th># DOCS</th>
<th>KEEP ▼</th>
<th>WEIGHT</th>
<th>ROLE</th>
<th>ATTRIBUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>credit card</td>
<td>5213</td>
<td>2874</td>
<td>✔</td>
<td>0.0650</td>
<td>Noun Group</td>
<td>Alpha</td>
</tr>
<tr>
<td>credit</td>
<td>2076</td>
<td>1579</td>
<td>✔</td>
<td>0.044</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>payment</td>
<td>2103</td>
<td>1387</td>
<td>✔</td>
<td>0.0040</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>account</td>
<td>1444</td>
<td>1160</td>
<td>✔</td>
<td>0.065</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>want</td>
<td>1406</td>
<td>1114</td>
<td>✔</td>
<td>0.062</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>pay</td>
<td>1475</td>
<td>927</td>
<td>✔</td>
<td>0.093</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>know</td>
<td>1210</td>
<td>857</td>
<td>✔</td>
<td>0.121</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>activate</td>
<td>838</td>
<td>730</td>
<td>✔</td>
<td>0.058</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>bill</td>
<td>926</td>
<td>684</td>
<td>✔</td>
<td>0.075</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>contact</td>
<td>763</td>
<td>682</td>
<td>✔</td>
<td>0.118</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>charge</td>
<td>642</td>
<td>651</td>
<td>✔</td>
<td>0.014</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>interest</td>
<td>743</td>
<td>596</td>
<td>✔</td>
<td>0.098</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>charge</td>
<td>705</td>
<td>541</td>
<td>✔</td>
<td>0.085</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>receive</td>
<td>612</td>
<td>503</td>
<td>✔</td>
<td>0.018</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>rate</td>
<td>621</td>
<td>498</td>
<td>✔</td>
<td>0.097</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>balance</td>
<td>614</td>
<td>406</td>
<td>✔</td>
<td>0.011</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
</tbody>
</table>
TEXT PARSING

- Documents are represented internally in SAS® Text Miner by a vector that contains the frequency of how many times each term occurs in each document.
TEXT PARSING

STEMMING

PART OF SPEECH

- Determines if the word is a common noun, verb, adjective, proper noun, adverb, etc.
- Disambiguate parts of speech when a word is used in a different context,
  - *I wish that my bank did not have a service charge for using other vendor ATM’s.*
  - *You can bank on either Germany or England winning the world cup next year.*

ENTITY EXTRACTION

<table>
<thead>
<tr>
<th>Places</th>
<th>People’s Names</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>White House</td>
<td>James H. Goodnight</td>
<td></td>
</tr>
</tbody>
</table>
PARTS OF SPEECH IN SAS® TEXT MINER

- Abbr (abbreviation)
- Adj (adjective)
- Adv (adverb)
- Aux (auxiliary or modal)
- Conj (conjunction)
- Det (determiner)
- Interj (interjection)
- Noun (noun)
- Num (number or numeric expression)

- Part (infinitive marker, negative participle, or possessive marker)
- Pref (prefix)
- Prep (preposition)
- Pron (pronoun)
- Prop (proper noun)
- Punct (punctuation)
- Verb (verb)
- VerbAdj (verb adjective)
ENTITIES IN SAS®
TEXT MINER

STANDARD ENTITIES (IDENTIFIED OUT-OF-THE-BOX)

- Address
- Company
- Currency
- Date
- Internet
- Location
- Measure
- Organization
- Percent

- Person
- Phone
- Prop_Misc (proper noun – ambiguous classification)
- SSN (U. S. Social Security Number)
- Time
- Time_Period
- Title
- Vehicle (motor vehicle)
ADDITI0NAL PARSING STEPS

• Specify Start/Stop/Synonym Lists
  • Filtering out low information words such as
    • articles (e.g. the, a, this)
    • prepositions (e.g. of, from, by)
    • conjunctions (e.g. and, but, or)
  • Consider document subject matter as well as domain-specific language and acronyms

• Vertical dictionaries
  • Automatically generate synonyms appropriate to the data
• Remove “boilerplate” language common to most or all documents
  • Headers and footers
  • Common qualifiers
  • Disclaimers
• Parse created data
  • Convert abbreviations
  • Correct misspellings
• Use term frequency filtering to assist with the creation of a stop list
• Recommendation: create subsets of documents by language. For example, all English documents in one corpus, all German documents in another corpus, etc.
• SAS includes extremely robust and sophisticated data manipulation capabilities, including character functions and regular expressions.
TEXT MINING PROCESS

- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis
• Also referred to as “Dimension Reduction”
• Transforms the quantitative representation into a compact and informative format
• Can also be used to further refine the data to be analyzed. For example, you can reduce the total number of parsed terms or documents that are analyzed.
• Eliminates extraneous information so that only the most valuable information or information that relates to a particular area of interest is considered.
• Spell checking
• Concept Linking
• Full text search
• Define additional synonyms
• Sub-setting management of terms and documents that are passed to subsequent nodes
• Singular value decomposition (SVD)

• Roll up terms

• Combination of both approaches
TEXT MINING PROCESS

- Text Preprocessing
- Text Parsing
- Transformation (Dimension Reduction)
- Document Analysis
• Expectation Maximization Clustering
  • Generates groups of similar documents from output of SVD
  • Fast clustering of many documents

• Hierarchical Clustering
  • Great for creating document taxonomies
• Note: each document is assigned to a single cluster

• Optionally, use unsupervised data mining methods like self organizing maps or clustering after building text mining clusters, using the text mining cluster segment identifiers as inputs in the subsequent analysis
• Discovers topics in document collection
• Allows automatic creation of single and multi-word topics
• User defined topics and editing of automatic topics
• Multiple topics per document
  • Soft clustering using rotated SVD (PROC SVD followed by PROC FACTOR)
Start with a table that contains either:
- Documents saved as a variable (column)
- A column that points to physical text files
EXAMPLE TEXT MINING PROCESS FLOWS

Apply natural language processing algorithms to parse the documents and quantify information about the terms in the corpus.
- Determine parts of speech (noun, verb, etc.)
- Perform stemming (run, runs, running, ran, etc.)
- Identify entities (names, places, etc.)
EXAMPLE TEXT MINING PROCESS FLOW

Optionally, filter the terms or documents that will be analyzed.
Can also perform spell-checking, full text searches, and analyze and view with Concept Linking
Analyze the documents to **create topics** and assign each document to one or more topics. In addition to derived topics, users can add their own topic definitions.
EXAMPLE DATA AND TEXT MINING PROCESS FLOW

Mining Structured Data

Mining Unstructured (Text) Data
DATA AND TEXT MINING PROCESS FLOW

Mining *ALL* Data: either Structured, or Unstructured or Both
The **Text Rule Builder** node generates an ordered set of rules that together are useful in describing and predicting a target variable.

Each rule in the set is associated with a specific target category, consisting of a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)).

A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3.
A tool providing a supervised approach to discovering and reporting the terms that best profile a set of documents associated with each level of a target variable.

- Uses a “new” procedure, Proc TMBelief, to determine the descriptive terms.
- Useful for binary, nominal, ordinal and date target variables.
- Internally we bin date variables to day, month, year etc. and map to ordinal.
- Note: User can bin interval target variables and then analyze as nominal or ordinal.
NEW NODE  TEXT PROFILE NODE

• How are men’s and women’s attitudes different toward my product?
• How has the answer to survey question #5 varied over the last 4 years?
• What is going on in the twitter feed over the last few months?
• Is there a difference in what people are talking about in different regions of the country?
• Visit
  http://support.sas.com/documentation/onlinedoc/txtminer/index.html
• Download “Getting Started with SAS Text Miner” (How to Guide) (Available for multiple versions)
• Download “Getting Started Examples (Zip)”
• Work to complete the examples.
SAS Text Miner Product Web Site


SAS Text Miner Technical Support Web Site
http://support.sas.com/software/products/txtminer/index.html

SAS Text Miner Technical Forum (Join Today!)

Data Mining and Text Mining Community

SAS Training

Data Miner Training Path:
http://support.sas.com/training/us/paths/dm.html

Courses for SAS® Text Miner:
YOUTUBE VIDEOS

• SASSoftware YouTube Channel
  • http://www.youtube.com/user/SASsoftware?feature=watch

• Manage All Unstructured Data with SAS® Text Analytics
  • http://www.youtube.com/watch?v=NHAq8jG4FX4&list=PL8BD07CC2C164FC40&index=4&feature=plpp_video

• SAS® Text Analytics Software Demo
  • http://www.youtube.com/watch?v=l1rYdrRCZJ4&feature=BFa&list=PL8BD07CC2C164FC40
QUESTIONS?
Thank you for your time and attention!

Connect with me:
LinkedIn: https://www.linkedin.com/in/melodierush
Twitter: @Melodie_Rush

CUSTOMER LOYALTY TEAM • Support You Can Count On