The SAS Text Mining Solution

Bernd.Drewes@eur.sas.com
Senior Data Mining Consultant
SAS International

Wayne.Thompson@sas.com
Product Manager, Enterprise Miner
SAS World Wide Marketing
Overview

- This presentation provides an overview of the SAS text mining solution including example applications.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or Trademarks of their respective companies.
Text Mining Timelines

- **Experimental**
  - Enterprise Miner 4.1, SAS 8.2, currently available
  - Early Adopter Text Mining Program

- **Production**
  - Add-on to Enterprise Miner 4.2 and 5.0, SAS 9.0, 1st quarter of 2002.
  - Separate API to potentially include text mining into other SAS products and solutions.
Early Adopter Program

Please note:

Enterprise Miner for Text is experimental and has a separate setinit. It will be made available on a restricted basis under an Early Adopter program. Please turn to your local SAS office for details."
Why Text Mining?
What is Text Mining?

- Discovering and using knowledge that exists in the document collection as a whole
  - Uncovering patterns within the collection
  - Establishing connections between documents and the terms in the collection as a whole
  - Combining free-form text and quantitative variables to derive information
Text Mining is not

- Natural Language Processing/Understanding
- Information Retrieval
- Knowledge extraction from a single document
Text Mining Applications

- Automatic Classification of Documents
  - Email filtering
  - Organize documents by topic into categories
  - News item routing
  - Generate classification codes from purchase descriptions

- Clustering
  - Survey data analysis
  - Customer complaints/comments
Text Mining Applications (cont.)

- Other forms of Prediction
  - Stock market prices from business news announcements
  - Customer satisfaction from customer comments
  - Cost based on call center log
How Does It Work?

- Document pre-processing
  - Data cleansing
  - Data reduction

- Document analysis
  - classification
  - clustering
  - prediction
  etc.
Leveraging Enterprise Miner
Data Input

- As a variable contained in a SAS data set.

- As documents residing on the file system.
Text Parsing (cont.)

- Stop lists
  - Filter out low information words
    - articles (e.g. the, a, this)
    - prepositions (e.g. of, from, by)
    - conjunctions (e.g. and, but, or)
    - etc.
  - Consider document subject matter
Text Parsing

- Create Term-Document Frequency Matrix
  - Record number of times term i appears in document j
  - Uses compressed representation in order to handle large collections

```
<table>
<thead>
<tr>
<th><em>TERM</em></th>
<th>TERMNUM</th>
<th>DOCUMENT</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>efficient</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>cross</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>database</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>transaction</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>processing</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>between</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```
Term Frequencies

- Which terms should be stopped?

<table>
<thead>
<tr>
<th>TERM</th>
<th>KEY</th>
<th>FREQ</th>
<th>NUMDOCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>sas</td>
<td>4</td>
<td>756</td>
<td>240</td>
</tr>
<tr>
<td>data</td>
<td>64</td>
<td>465</td>
<td>151</td>
</tr>
<tr>
<td>paper</td>
<td>58</td>
<td>147</td>
<td>112</td>
</tr>
<tr>
<td>system</td>
<td>36</td>
<td>135</td>
<td>78</td>
</tr>
<tr>
<td>more</td>
<td>195</td>
<td>81</td>
<td>61</td>
</tr>
<tr>
<td>information</td>
<td>207</td>
<td>97</td>
<td>58</td>
</tr>
<tr>
<td>users</td>
<td>30</td>
<td>79</td>
<td>57</td>
</tr>
<tr>
<td>proc</td>
<td>148</td>
<td>145</td>
<td>56</td>
</tr>
<tr>
<td>application</td>
<td>165</td>
<td>80</td>
<td>48</td>
</tr>
<tr>
<td>base</td>
<td>91</td>
<td>49</td>
<td>47</td>
</tr>
<tr>
<td>time</td>
<td>119</td>
<td>67</td>
<td>47</td>
</tr>
<tr>
<td>applications</td>
<td>712</td>
<td>71</td>
<td>45</td>
</tr>
<tr>
<td>software</td>
<td>60</td>
<td>67</td>
<td>45</td>
</tr>
<tr>
<td>code</td>
<td>97</td>
<td>62</td>
<td>44</td>
</tr>
<tr>
<td>output</td>
<td>6</td>
<td>99</td>
<td>43</td>
</tr>
<tr>
<td>based</td>
<td>141</td>
<td>50</td>
<td>39</td>
</tr>
<tr>
<td>web</td>
<td>1125</td>
<td>91</td>
<td>39</td>
</tr>
<tr>
<td>macro</td>
<td>96</td>
<td>72</td>
<td>37</td>
</tr>
<tr>
<td>analysis</td>
<td>574</td>
<td>60</td>
<td>36</td>
</tr>
<tr>
<td>access</td>
<td>35</td>
<td>43</td>
<td>35</td>
</tr>
</tbody>
</table>
Weighing words

- Weigh words differently. Actually done in the SVD node
  - focus on words that help to distinguish between texts
  - de-emphasize words that
    - are common to most documents
    - only occur rarely
Singular Value Decomposition (SVD)

- Dimension reduction
- From thousands to hundreds or less
- Terms and documents projected into the same space
- Documents with similar meaning (but dissimilar word patterns) can still be near one another
- Other variables can be combined with the results of the SVD
  - Quantitative variables
  - Categorical variables
Benefits of SVD

- Assumes that there is something conceptual underlying words (obscured by word choice)
- these concepts can be expressed as collections or combination of words
- SVD is used to estimate the structure of these concepts and find their most relevant words
Text Clustering (E-M Cluster)

- Generates groups of similar documents
- from output of SVD
- fast clustering of many documents
- each document may belong to several classes ("fuzzy clustering")
Mining the Text

- Utilize Enterprise Miner's tools
  - Memory-Based Reasoning
  - Neural Network
  - Regression
  - etc.

- Combine other variables with the results of the SVD
  - Quantitative variables
  - Categorical variables
  - Target variables
Review of the Text Mining Process

- Text parsing extracts terms and creates a term-document frequency matrix
- Each document is characterized by a vector in multi-dimensional space
- The vectors are then reduced down to 100 or so dimensions by the SVD
- Results can be classified, clustered or used for prediction with existing Enterprise Miner tools