Using SAS Text Miner to extract intelligence from unstructured text sources

Manya Mayes
SAS Text Miner Product Manager

Bernd Drewes
SAS EMEA Text Miner Product Manager
What is Text Mining?
Text Miner and the SAS Intelligence Value Chain

- **SAS Capabilities**
  - Comprehensive text and data mining

- **Business Value**
  - Reduce cycle times
SAS Text Miner

- Text pre-processing
- Transformation and reduction
- Text clustering
- Concept linking
- Fully integrated with Enterprise Miner
Text Miner Interactive Results Browser

7,116 Documents

<table>
<thead>
<tr>
<th>Term</th>
<th>Freq</th>
<th># Documents</th>
<th>Keep</th>
<th>Links</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>4128</td>
<td>3019</td>
<td>Y</td>
<td>0</td>
<td>0.111</td>
</tr>
<tr>
<td>not</td>
<td>3027</td>
<td>2205</td>
<td>Y</td>
<td>0</td>
<td>0.143</td>
</tr>
<tr>
<td>brake</td>
<td>3006</td>
<td>2201</td>
<td>Y</td>
<td>32</td>
<td>0.153</td>
</tr>
<tr>
<td>on</td>
<td>2689</td>
<td>1885</td>
<td>Y</td>
<td>0</td>
<td>0.166</td>
</tr>
<tr>
<td>accident</td>
<td>2005</td>
<td>1617</td>
<td>Y</td>
<td>10</td>
<td>0.160</td>
</tr>
<tr>
<td>cause</td>
<td>1667</td>
<td>1379</td>
<td>Y</td>
<td>11</td>
<td>0.162</td>
</tr>
<tr>
<td>do</td>
<td>1500</td>
<td>1928</td>
<td>Y</td>
<td>0</td>
<td>0.174</td>
</tr>
<tr>
<td>driver</td>
<td>1838</td>
<td>1498</td>
<td>Y</td>
<td>33</td>
<td>0.184</td>
</tr>
<tr>
<td>deploy</td>
<td>1463</td>
<td>1389</td>
<td>Y</td>
<td>30</td>
<td>0.187</td>
</tr>
<tr>
<td>with</td>
<td>1239</td>
<td>1192</td>
<td>Y</td>
<td>29</td>
<td>0.203</td>
</tr>
<tr>
<td>hit</td>
<td>1320</td>
<td>1174</td>
<td>Y</td>
<td>16</td>
<td>0.200</td>
</tr>
<tr>
<td>in</td>
<td>1331</td>
<td>1149</td>
<td>Y</td>
<td>0</td>
<td>0.212</td>
</tr>
<tr>
<td>when</td>
<td>1265</td>
<td>1145</td>
<td>Y</td>
<td>0</td>
<td>0.211</td>
</tr>
<tr>
<td>consumer</td>
<td>1587</td>
<td>1021</td>
<td>Y</td>
<td>19</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Clusters

<table>
<thead>
<tr>
<th>#</th>
<th>Descriptive Terms</th>
<th>Freq</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>+ impact, +lock, position, up, +brake</td>
<td>337</td>
<td>0.0473586563</td>
</tr>
<tr>
<td>3</td>
<td>+control, +lose, +roll, over, +road</td>
<td>438</td>
<td>0.061514334</td>
</tr>
<tr>
<td>4</td>
<td>+brake, +stop, +decel, info, +consumer</td>
<td>1221</td>
<td>0.1715851602</td>
</tr>
<tr>
<td>5</td>
<td>+jump, +transmission, out of, +roll, +park</td>
<td>430</td>
<td>0.059154134</td>
</tr>
<tr>
<td>6</td>
<td>+car, rear, on, +hit, when</td>
<td>430</td>
<td>0.059154134</td>
</tr>
<tr>
<td>7</td>
<td>+seat, +distracted, +driver, +passenger, +injury</td>
<td>553</td>
<td>0.0777212972</td>
</tr>
<tr>
<td>8</td>
<td>+driver, +do, +collision, mph, +air</td>
<td>721</td>
<td>0.1013288861</td>
</tr>
<tr>
<td>9</td>
<td>+axle, +wheel, off, rear, +break</td>
<td>522</td>
<td>0.0739558179</td>
</tr>
<tr>
<td>10</td>
<td>+result, abs, +failure, +accident, +system</td>
<td>317</td>
<td>0.0445474038</td>
</tr>
<tr>
<td>11</td>
<td>+tie, +number, +tread, dot number, real</td>
<td>378</td>
<td>0.0531197302</td>
</tr>
<tr>
<td>12</td>
<td>+accelerate, causing accident, acceleration, +result, in</td>
<td>300</td>
<td>0.0427175937</td>
</tr>
<tr>
<td>13</td>
<td>+airbag, +do, ramp, +airbag, +driver</td>
<td>521</td>
<td>0.0732152808</td>
</tr>
<tr>
<td>14</td>
<td>during, no, above, +list, summary</td>
<td>114</td>
<td>0.0160202381</td>
</tr>
</tbody>
</table>
New to SAS Text Miner 9.1

- additional language support
- Web crawling
- additional options for text parsing
- new functionalities in Text Miner Results Browser including new graphics
- improved performance
- specific error messages
- And more…
Text Miner Concept Linking
Text Miner Hierarchical Clustering Browser
Text Mining Applications

- Early Warning
- SKU Classification
- Call center classification
- Pharma literature support
- Extracting information from police reports
Industrial Applications

- Early Warning
- SKU Classification
- Call Center Classification

For more detail see ppt’s/talks by
  - Randy Collica: Wed 11.35
  - Manya Mayes & Randy Collica: Today, 10.50
Emerging applications: Extracting knowledge from Text

- Pharma & Life Sciences
  - Process predefined taxonomies
  - Classify texts into taxonomy
  - Learn classification rules
  - Cluster by content

- Police/Government/Intelligence
  - Extract content descriptors
  - Classify into classes & hierarchies
  - Match & link texts
Scientific Abstracts in Pharma

Genetic polymorphisms in drug-metabolizing enzymes are responsible for interindividual differences in the disposition of numerous drugs. The poor metabolizers (PMs) are derived from mutant alleles of cytochrome P450 (CYP), and those PMs are more susceptible to the development of adverse drug reactions, when the parent drug is pharmacologically active, or are showing a lack of therapeutic efficacy, when a prodrug is administered and its active metabolite(s) is not formed. This article reviews the basic knowledge of genetic polymorphisms of CYP-related drug-metabolizing enzymes and the clinical implications for a safer and more efficient drug therapy.
The poor metabolizers (PMs) are derived from mutant alleles of cytochrome P450 (CYP), and those PMs are more susceptible to the development of adverse drug reactions... This article reviews the basic knowledge of genetic polymorphisms of CYP-related drug-metabolizing enzymes.
Text Mining Tasks

- Highly technical abstracts - Need to:
  - generate domain terminology
  - Process domain content structure
    - Gene ontology
  - Apply predictive text modeling
  - Learn classification rules
  - Cluster by content
Cytochrome P450 Gene Ontology (subset)

- GO:0008403 : 25-hydroxycholecalciferol-24-hydroxylase (0)
- GO:0008391 : arachidonic acid monooxygenase
- GO:0008402 : aromatase (1)
- GO:0004498 : calcidiol 1-monooxygenase (2)
- GO:0008123 : cholesterol 7-alpha-monooxygenase (0)
- GO:0008389 : coumarin 7-hydroxylase (2)
- GO:0004500 : dopamine-beta-monooxygenase
Text Mining

1. Texts
2. Text Partition
3. Text Miner with domain indices
4. Tree
5. Find neighbors of domain indices
6. Text Miner with open vocabulary
7. Text Partition
8. Text Miner
9. Tree
10. TEXTS with multiple classes
11. Assess correctness
Learning Classification Rules
## Text Clusters

<table>
<thead>
<tr>
<th>Descriptive Terms</th>
<th>Frequency</th>
<th>Proportion of Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>monooxygenase, cytochrome, monooxygenase activities</td>
<td>98</td>
<td>0.081803005</td>
</tr>
<tr>
<td>hydroxylase activity, hydroxylase, testosterone, lauric, arachidonic</td>
<td>245</td>
<td>0.2045075125</td>
</tr>
<tr>
<td>fatty acids, fatty, nitric, synthase, prostaglandin</td>
<td>185</td>
<td>0.1544240401</td>
</tr>
<tr>
<td>cytochrome, reductase</td>
<td>246</td>
<td>0.2053422371</td>
</tr>
<tr>
<td>cholesterol, coumarin, specific, sterol, demethylase</td>
<td>220</td>
<td>0.183639395</td>
</tr>
<tr>
<td>aromatase activity, aromatase, beta, specific, testosterone</td>
<td>204</td>
<td>0.1702838063</td>
</tr>
</tbody>
</table>
From Pharma to Police

- Domains are very different
  - Police: ordinary language
    - more support available

- Tasks in many ways similar
  - Provide content description
  - Classify, cluster, link texts
Text Mining of Police Reports

Goals

- Extract Content
  - Automatically annotate texts with descriptors
  - Classify texts
  - Cluster similar texts
  - Link texts through common descriptors
- Eventually: Model Content, such as
  - Find most important predictors for crimes
    - Of a certain type
    - In a certain area, etc.
Example: Police Report

Former Walt Whitman High School student Ryan Howard Baird pleaded “involved” – the juvenile equivalent of “guilty” – today in the juvenile division of the Montgomery County Circuit Court for sexually assaulting and robbing a Gaitherburg woman in November.

Baird pleaded guilty to two counts of a first-degree sex offense, first-degree rape, robbery with a dangerous and deadly weapon, first-degree assault, a third-degree sex offense, and false imprisonment, Wennar said.
Project: Processing Police Reports

- Build descriptors for data mining purposes:
  - Generate Domain Specific Vocabulary with Text Miner
  - Group Vocabulary into domain categories (persons, locations, crimes, pleads,...)
  - Characterize text as table of domain entries

- Find central message by focusing on terms with highest degree of connections

- Next slide: SAS TM & Data Step found…
Some of the Descriptors found

<table>
<thead>
<tr>
<th>Person</th>
<th>Ryan Howard Baird</th>
<th>student</th>
<th>Juvenile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Walt Whitman High School</td>
<td>Wennar</td>
<td>Baird</td>
</tr>
<tr>
<td>Place</td>
<td>Montgomery County</td>
<td>Gaithersburg</td>
<td></td>
</tr>
<tr>
<td>Pleads</td>
<td>Involved</td>
<td>guilty</td>
<td></td>
</tr>
<tr>
<td>Crime</td>
<td>Sex</td>
<td>Assault</td>
<td>robbery</td>
</tr>
<tr>
<td>Terms</td>
<td>Dangerous</td>
<td>Deadly</td>
<td>weapon</td>
</tr>
<tr>
<td>Charge</td>
<td>First-degree sex offense</td>
<td>First-degree rape</td>
<td>First-degree assault</td>
</tr>
</tbody>
</table>
Annotate Texts Automatically

- Finding content descriptors
  - “conspiracy commit robbery”, “child abuse”, “first-degree murder”, ”third-degree sex offense

- Using descriptors
  - Content descriptors: For classifying & similarity groupings
  - Indices (names etc.): For interconnecting texts

- Classify into topics or hierarchies
  - Based on content descriptors
  - Optionally structured by human intervention
Make use of annotations: Classify and Interconnect

- Use each descriptor (triple, pair, single) as variable
- Compute for each text the occurrence of all such variables
- Use predictive modeling to assign texts to classes or hierarchies
- Inter-Connect texts with descriptors
Finally: Apply Data Mining for…

- **Root Cause Analysis**
  - Find characteristics of particular crimes
  - Find crime patterns for particular groups (e.g. juveniles in Montgomery)

- **Find Links**
  - Connections between groups of people that may be involved in distinct crimes
  - between current cases and historical data
    - Generate list suspects and/or profiles for current case

- **Find crime clusters with similar characteristics**
Summary

- SAS Text Mining Product Characteristics
- Successful Industrial Applications
- Emerging application
- Emerging architectures
  - Data mining & text mining
  - Mining & OLAP
  - Interactive Mining & Searching
Related Talks

- Topic Scoring Engine: Wed, 17.35, also today 13.30
- Annotation of Gene Products, Today 10.50