Alberta Parks’ use of Text Mining

Alberta Tourism, Parks and Recreation,
Parks Division,
Business Integration and Analysis
Statistics
Man!
realistically...
• 209 Provincial Recreation Areas
• 75 Provincial Parks
• And more...

Total 478

Alberta Park history: http://www.tpr.alberta.ca/parks/managing/history.asp
Transforming Data

Data is shapeless, providing limited insight without the proper tools to drive fact-based analysis and decision making.
Analytics Magic?
“Most organizations are drowning in data but starving for information.”

Anonymous
Data: Structured v.s. Unstructured

Three Major Areas within Data Continuum

Structured Data
- Relational databases
- Spreadsheets
- Flat files in record format
- Legacy databases: Hierarchical, Mainframe

Semi-structured Data
- XML Docs
- EDI Docs
- RSS Feeds
- Web Logs
- Word processing files
- Web Pages
- Multimedia

Unstructured Data
- E-mail
- Content Management
- Document Management
- Voice recognition
- Instant Messaging
- Taxonomies, Ontologies
- Wikis

Decreasing Moderately (-15% to -46%)
Increasing Moderately (18% to 47%)
Increasing Sharply (61% to 81%)

Anticipated Decrease or Increase over Next Three Years
Unstructured Data Sources

- Web page
- Email
- Content management system records
- Word Document, PDF
- Telephone call
- Instant message
- SMS (text message)
- Letters from the public
- Tweet from Twitter
- Blog post
- etc...
Text Analytics

Using statistical methods to analyze and interpret the meaning of **textual data** (unstructured data).

• Visionary paper written by Hans Peter Luhn titled “The Automatic Creation of Literature Abstracts” for the 1958 IBM Journal marks birth of computational text analytics

• Automated **solutions** go mainstream in **early 2000’s** by visionary companies such as SAS and Teragram

• **Web 2.0** has kicked off an arms race to **capture** the broad and vast **content** now being **exposed** by the web (I.e. Online social networks)
Text Analytics

Information Organization and Access

Predictive Modeling, Discover Trends and Patterns

Enterprise Content Categorization

Ontology Management

Text Mining

Sentiment Analysis
Text Analytics

Information Organization and Access
- Enterprise Content Categorization
- Ontology Management

Predictive Modeling, Discover Trends and Patterns
- Sentiment Analysis

Text Mining
Business Intelligence Platform

Social Media Analytics

Other Analytics
Text Mining
Natural Language Processing

- Stem...Stems...Stemming (park, parks, parking)
- Parts of Speech (verb, noun, adjective...)
- Dictionaries
- Entity Extraction

person    place    dates
Natural Language Processing

- UPPERCASE
- Miss-spelings
- A.C.R.O.N.Y.M.S
- Shrt-hnd
- Prof@nity
- *Punctuation*
Data Mining

Data Mining is applied statistics and pattern recognition to discover knowledge from data.

Any large number of observations and variables can be data mined for valuable information.
Text Mining

“The process of discovering and extracting meaningful patterns and relationships from text collections.”
TM = Discovery ≠ Search

• Text Mining is **more than frequency counts**. Frequency excludes context and relations.

• A Microsoft Word ‘word count’ or a word cloud does not capture meaning and could even be misleading.

• Text Mining helps discover key concepts, term associations and relationships.
Search v.s. Discovery

You can search the island, but you might still be lost
Is text really unstructured data?
Text Analytics Adoption Curve

UNDISCIPLINED
- Content tagging and retrospective indexing
- Manual review of web content and emails for customer insight
- Largely manual, redundant and error-prone

REACTIVE
- Automatic content categorization and taxonomy management
- Summarization of existing materials
- Largely focused in internal collection organization, increase efficiency and standardization

PROACTIVE
- Extract sentiments, facts for trend monitoring
- Mine unstructured/semi-structured data for previously unknown patterns
- Discover previously unknown patterns, create new insights and knowledge that inform and automate business processes and increase the reliability of predictive and descriptive models

GOVERNED
- Create enterprise ontology for unstructured and semi-structured data
- Unify metadata management to integrate structured data with contextual analysis

Value
- Low
- High

Risk
- Low
- High

People, Policies, Technology Adoption
Some things in parks can’t be analysed...
“What could we have done to make your visit better?”
- Annual camper satisfaction survey
“The nice looking lady that woke me up was a very good start”
“Great skinny dipping lake, nothing wrong there.”
“A great big Budweiser motorhome pulls up with 12 girls that want to party!”
“Fireworks at 4:00 in the morning”
“Daughter says more hot boys are needed”
“Park Rangers were informative – maybe it was just my pretty girlfriend”
“Conservation officer was a hottie. He should have visited more often”
“We are two spry young gents and enjoy camping here.”
“One ply toilet paper is insufficient for the task”
“Rangers need skimpier uniforms”
“Well, you could have stopped by for a little cuddle time...”
• Hand written responses are not ideal. It comes with the nature (no pun intended) of our business.

• Comments are transferred to electronic format through typing or Speech to Text (dictation) software.
Old method – Assigning codes

187 sub-categories across 28 General Categories.
Examples:

- Washrooms
- Firewood
- Roads
- Showers
- Security
- Grounds Maintenance
- Operations

- Information Services
- Pest Control
- Playgrounds
- Reservation System
- Fishing
- Facilities
- Beach / Lake

- Policy
- Trails
- Camping Preferences
- Value
- Noise
Once comments are assigned codes, simple frequency counts show magnitudes of customer feedback...

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>code_1</td>
<td>code_2</td>
<td>code_3</td>
<td>code_4</td>
<td>code_5</td>
</tr>
<tr>
<td>2</td>
<td>16b</td>
<td>16b</td>
<td>23f</td>
<td>16b</td>
<td>16b</td>
</tr>
<tr>
<td>3</td>
<td>16b</td>
<td>23f</td>
<td>16b</td>
<td>100m</td>
<td>100b</td>
</tr>
<tr>
<td>4</td>
<td>100a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>100a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8a</td>
<td>8a</td>
<td>8a</td>
<td>8a</td>
<td>16b</td>
</tr>
<tr>
<td>7</td>
<td>15b</td>
<td>3d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>100c</td>
<td>15b</td>
<td>16b</td>
<td>23f</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4a</td>
<td>4a</td>
<td>2f</td>
<td>5a</td>
<td>100b</td>
</tr>
<tr>
<td>12</td>
<td>4a</td>
<td>3d</td>
<td>5a</td>
<td>16b</td>
<td>16e</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...see example of frequency counts on the next slide...
## 2008 Camper Satisfaction Survey
### General and Sub-Category Comments - Provincial Negative Comments
*(Total Surveys Represented - 1,118)*

<table>
<thead>
<tr>
<th>General Category</th>
<th>Sub-Category</th>
<th># of Comments</th>
<th>% of Category</th>
<th>% of All Comments</th>
<th>% of ALL Surveys Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firewood</strong></td>
<td>Too expensive</td>
<td>96</td>
<td>30.1</td>
<td>3.6</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>Firewood Quantity (not enough/no wood)</td>
<td>61</td>
<td>19.1</td>
<td>2.3</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>Poor Quality (too long, wet)</td>
<td>48</td>
<td>15.0</td>
<td>1.8</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Poor Access (location, timing)</td>
<td>47</td>
<td>14.7</td>
<td>1.8</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Should be free</td>
<td>40</td>
<td>12.5</td>
<td>1.5</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>Firewood Delivery Needed and other</td>
<td>13</td>
<td>4.1</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Firewood Should be Included in Fees</td>
<td>12</td>
<td>3.8</td>
<td>0.4</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Firewood Shelter Needed/Upgraded</td>
<td>2</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td></td>
<td>319</td>
<td>100.0</td>
<td>11.9</td>
<td>28.5</td>
</tr>
<tr>
<td><strong>Hook-ups/Dump stations/Water</strong></td>
<td>Additional power campsites</td>
<td>86</td>
<td>34.8</td>
<td>3.2</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>Full Power-Water-Sewer Hook-ups Needed</td>
<td>31</td>
<td>12.6</td>
<td>1.2</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>Other (specific amperage, water filling station needed)</td>
<td>26</td>
<td>10.5</td>
<td>1.0</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>More Taps / Water Locations</td>
<td>24</td>
<td>9.7</td>
<td>0.9</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Poor Drinking Water Quality / Need Potable Water</td>
<td>21</td>
<td>8.5</td>
<td>0.8</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Install power campsites</td>
<td>20</td>
<td>8.1</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Sewage Dump-stations Needed / Dirty / Full</td>
<td>18</td>
<td>7.3</td>
<td>0.7</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Water Hook-ups Needed</td>
<td>11</td>
<td>4.5</td>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Running Water Needed (not washroom related)</td>
<td>10</td>
<td>4.0</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td></td>
<td>247</td>
<td>100.0</td>
<td>9.2</td>
<td>22.1</td>
</tr>
</tbody>
</table>
Analysing Text

The Old Way
1. Typing comments (~3 weeks/year)
2. Every comment manually read and manually assigned special codes (~3 weeks/year)

The New Way
1. Dictation software types comments (~1 week/year)
2. SAS Text Miner analyses data (~1 minute/year)*

*First year requires a few days to create the ‘black box’ but becomes a production run thereafter.
## 2008 Camper Satisfaction Survey
### General and Sub-Category Comments - Provincial Negative Comments
*(Total Surveys Represented - 1,118)*

<table>
<thead>
<tr>
<th>General Category</th>
<th>Sub-Category</th>
<th># of Comments</th>
<th>% of Category</th>
<th>% of All Comments</th>
<th>% of ALL Surveys Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firewood</td>
<td>Too expensive</td>
<td>96</td>
<td>30.1</td>
<td>3.6</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>Firewood Quantity (not enough/no wood)</td>
<td>61</td>
<td>19.1</td>
<td>2.3</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>Poor Quality (too long, wet)</td>
<td>48</td>
<td>15.0</td>
<td>1.8</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Poor Access (location, timing)</td>
<td>47</td>
<td>14.7</td>
<td>1.8</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Should be free</td>
<td>40</td>
<td>12.5</td>
<td>1.5</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>Firewood Delivery Needed and other</td>
<td>13</td>
<td>4.1</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Firewood Should be Included in Fees</td>
<td>12</td>
<td>3.8</td>
<td>0.4</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Firewood Shelter Needed/Upgraded</td>
<td>2</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal</strong></td>
<td><strong>319</strong></td>
<td><strong>100.0</strong></td>
<td><strong>11.9</strong></td>
<td><strong>28.5</strong></td>
</tr>
<tr>
<td>Hook-ups/Dump stations/Water</td>
<td>Additional power campsites</td>
<td>86</td>
<td>34.8</td>
<td>3.2</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>Full Power-Water-Sewer Hook-ups Needed</td>
<td>31</td>
<td>12.6</td>
<td>1.2</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>Other (specific amperage, water filling station needed)</td>
<td>26</td>
<td>10.5</td>
<td>1.0</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>More Taps / Water Locations</td>
<td>24</td>
<td>9.7</td>
<td>0.9</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Poor Drinking Water Quality / Need Potable Water</td>
<td>21</td>
<td>8.5</td>
<td>0.8</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Install power campsites</td>
<td>20</td>
<td>8.1</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Sewage Dump-stations Needed / Dirty / Full</td>
<td>18</td>
<td>7.3</td>
<td>0.7</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Water Hook-ups Needed</td>
<td>11</td>
<td>4.5</td>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Running Water Needed (not washroom related)</td>
<td>10</td>
<td>4.0</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal</strong></td>
<td><strong>247</strong></td>
<td><strong>100.0</strong></td>
<td><strong>9.2</strong></td>
<td><strong>22.1</strong></td>
</tr>
</tbody>
</table>
Using SAS Text Miner...

firewood (234) + wood (100) = 334 (v.s. 319)
% Represented = 25% (v.s. 28.5%)
Firewood’s related terms ("sub-categories")
Refining the model...

Synonyms:

- firewood = wood
- include = bundle
- and more...

Text Miner’s ability to set synonyms and handle other lexical relations outweighs and outperforms days spent re-categorizing.
Refining the model...

Ability to handle synonyms
Continuous refining

- + price
- + bring
- free firewood
- + cheap
- + dry
- + fire
- + include
- expensive
- free
Actionable Intelligence

There is no way to determine outcomes like this:

“Percieved camper safety can be impacted by the level of noise, bathroom or site cleanliness, and the amount of officer patrols. Failing in any of these may contribute to campers feeling unsafe”.

From output like this:
Noise in parks?
Noise in parks

Leveraging Existing Data

• Survey comments 2002 – 2011
• 18,510 comments

Analysis

• Text Mining
• No specific park stood out as a problem area
  ➔ Not to say there are none
  ➔ Only those ~100 top visited parks
Noise in Parks? - Results

*Noise is 6% of all comments*

- Generators = 1.8%
- Parties = 1.3%
- Music = 1.2%
- Barking = 0.7%

*some double counting*

The magnitude of the problem is no bigger than other problems (e.g. Boat launch, road issues), but the sentiment is strong, making this an important issue.
Noise in Parks? - Sentiment

*Sentiment:*

- 10% specifically mentioned banning generators.
- The remainder demand quiet time respect.
- More patrols to better control noise.
- A few suggested identifying sites for generator users.

- Educating (improve information services) and improved enforcement are suggested.
Letters from the Public
(a.k.a. Action Requests)
Action Requests – Text Mining

Difficult because AR process is not built with the mindset that public feedback is data.

Dataset used in this example is a folder of PDF documents painstakingly downloaded from ARTS, one click at a time.

(Courtesy of Peter Weclaw ☺️)
Action Requests – Bill 29 dialogue

Each PDF contains public letters AND our response

Example of TM’s accuracy and Clustering feature:

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Descriptive Terms</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>'ecological integrity', integrity, ecological</td>
<td>...</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>conservation activities, land</td>
<td>...</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>public activities proposed</td>
<td>...</td>
<td>42</td>
</tr>
</tbody>
</table>
Action Requests – Bill 29
Social Media Monitoring

yammer

twitter
Who’s on Yammer?
Information is the new Currency $$$
New Website ROI:
www.AlbertaParks.ca
Old v.s. New AlbertaParks.ca website

Comparing consistency and findability of information of the new website v.s. the old website.
Email as a data source
Spam-a-thon

• One mild mannered Wednesday, GOA was hit with massive amounts of Spam email.

• Recipients of the spam kept responding to the spam and everyone on the list would get it. (i.e. We were spamming ourselves).

• 138 employees responded to the spam → nearly 1 email every 2 minutes.
Spam-a-thon

• Demonstration of analytics and visualization of the SPAM event

• The insight provides a new perspective on the problem and educates people on a better understanding of spam issues.

• EMAIL as a data source
Spam-a-thon

• 7% of people marked the spam with high importance.
• Half (50%) asked to be removed from the distribution list
• 27% asked the spammer to ‘just stop’.
• Our Canadian side shines brightly since 54% of all responses contained Please and/or Thank you...
Spam-a-thon: Canadian Politeness
Spam-a-thon

• 14% of people replied with "Ditto“ (E.g. they write "Me too“)

• 70% of the responses came from women (I’m not going to read into this, but feel free to make your own interpretations with your work pals)

• 20% of people responded to tell people that by responding, they are contributing to the spam. This perpetuated the problem. Oh, the irony.
Graphing length of response text
“Nothing is more terrible than activity without insight.”

Thomas Carlyle
Contact Information

Jared Prins B.Sc., Program Analyst

Business Integration and Analysis Section

Alberta Tourism, Parks & Recreation

2nd Flr. Oxbridge Place, 9820-106 Street

Edmonton, AB

T5K 2J6

Phone: 780.427.6313

Fax: 780.427.5980

www.AlbertaParks.ca