AI vs. Automation

Workshop
<table>
<thead>
<tr>
<th>Length</th>
<th>Format</th>
<th>Description</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>Activity Demos</td>
<td>▪ Demo some natural language processing, entity extraction, and complex Boolean query tagging tools (Lexalytics, Data Harmony, IBM Watson)</td>
<td>▪ Understand the types of features that differentiate automated tagging tools, and the criteria for evaluating them.</td>
</tr>
<tr>
<td>50 min</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 min</td>
<td>Lecture Activity</td>
<td>▪ Participate in a query building exercise.</td>
<td>▪ Obtain a practical understanding of how to build an automated classifier.</td>
</tr>
<tr>
<td>25 min</td>
<td>Show/Tell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 min</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Who’s in the room?

1) Pick the **one** profession you most strongly identify with.

<table>
<thead>
<tr>
<th>Profession</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librarian/Archivists</td>
<td>9</td>
</tr>
<tr>
<td>Taxonomist</td>
<td>2</td>
</tr>
<tr>
<td>Info/Data Scientist</td>
<td>2</td>
</tr>
<tr>
<td>Researcher</td>
<td>1</td>
</tr>
<tr>
<td>Student</td>
<td>0</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>1</td>
</tr>
<tr>
<td>SW Engineer</td>
<td>1</td>
</tr>
<tr>
<td>Information Architect</td>
<td>1</td>
</tr>
<tr>
<td>Other Interested Party</td>
<td>1</td>
</tr>
</tbody>
</table>
Outline

- Tool demos
  - What are Boolean queries, how to build them, and why
## Tool Demos

<table>
<thead>
<tr>
<th>Tool</th>
<th>Demo URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Harmony *</td>
<td><a href="http://demo.newsindexer.com/">http://demo.newsindexer.com/</a></td>
</tr>
<tr>
<td>IBM Watson</td>
<td><a href="https://natural-language-understanding-demo.ng.bluemix.net/">https://natural-language-understanding-demo.ng.bluemix.net/</a></td>
</tr>
<tr>
<td>Intellexer</td>
<td><a href="http://demo.intellexer.com/">http://demo.intellexer.com/</a></td>
</tr>
<tr>
<td>Lexalytics *</td>
<td><a href="https://www.lexalytics.com/demo">https://www.lexalytics.com/demo</a></td>
</tr>
<tr>
<td>Meaning Cloud</td>
<td><a href="https://www.meaningcloud.com/demo">https://www.meaningcloud.com/demo</a></td>
</tr>
<tr>
<td>PoolParty PowerTagging</td>
<td><a href="https://drupal.poolparty.biz/powertagging">https://drupal.poolparty.biz/powertagging</a></td>
</tr>
<tr>
<td>Text Razor</td>
<td><a href="https://www.textrazor.com/demo">https://www.textrazor.com/demo</a></td>
</tr>
</tbody>
</table>

* In-depth demos using web client applications for these tools.
Outline

- Tool demos
- What are Boolean queries, how to build them, and why
Case study

Lexalytics

Childhood Obesity  Disease Prevention and Health Promotion  Health Care Quality  Health Coverage

Robert Wood Johnson Foundation
## Testing process: Text collections

<table>
<thead>
<tr>
<th>User interfaces</th>
<th>API</th>
<th>Excel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test collection</td>
<td>90 Repository long form assets</td>
<td>400 WCMS short form assets</td>
</tr>
<tr>
<td>Content</td>
<td>Full text</td>
<td>Title &amp; summary only</td>
</tr>
<tr>
<td>Format</td>
<td>Clear text</td>
<td>Clear text, CSV</td>
</tr>
<tr>
<td>Topics</td>
<td>Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage</td>
<td>Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage</td>
</tr>
</tbody>
</table>
Test process: Categorization (to a pre-defined set of categories)

- Build and test a rule
  - A Boolean query with proximity operators to classify into a Topic (called a “configuration” in Lexalytics Semantria).
- Modify and test a rule.
- Obtain relevant classification
  - Identify the correct Topic, 80% or more of the time.
  - If an incorrect Topic is returned, why was it returned? Is an incorrect Topic potentially relevant?
Sample pre-defined query: Health coverage

("health insurance coverage" OR "health coverage") OR ("healthcare reform" OR "health care reform") OR ("Better Care Reconciliation Act" OR BCRA) OR ("American Health Care Act" OR AHCA) OR ("Affordable Care Act" OR "ACA" OR Obamacare) OR ((Medicare OR Medicaid) NEAR/5 (spend* OR cover* OR expan*)) OR ("health insurance exchange" OR "HIE") OR ("health insurance" NEAR/5 marketplace*) OR ("federal* facilitated marketplace**" NEAR/10 "health insurance") OR ("federal* run marketplace**" NEAR/10 "health insurance") OR ((state NEAR/5 marketplace*) NEAR/10 "health insurance") OR ("small business marketplace**" NEAR/10 "health insurance") OR ("small-business marketplace**" NEAR/10 "health insurance") OR ("small business" NEAR/5 exchange*) NEAR/10 "health insurance") OR ("high-risk" OR "high risk") NEAR/10 "health insurance") OR (uninsured NEAR/5 (veteran* OR child* OR adult* OR people OR kid* OR citizen*)) OR ("pre-existing condition**" OR "preexisting condition**") NEAR/10 "health insurance") OR "health insurance rate**" OR ((cost* OR rate* OR payment*) NEAR/10 "health insurance") OR ("health insurance" NEAR/10 "tax credit") OR (healthcare OR "health care") NEAR/5 spending) OR (healthcare OR "health care") NEAR/5 utilization) OR ("high-deductible" OR "high deductible") NEAR/10 "health insurance") OR ("mental health" OR "substance abuse") NEAR/10 "health insurance") OR ("provider network**" NEAR/10 "health insurance") OR ("in-network" OR "out-of-network") NEAR/10 "health insurance") OR (("PPO* OR HMO") NEAR/5 (marketplace* OR plan* OR provider*)) OR ("health insurance" NEAR/10 (enroll* OR "re-enroll**" OR renew* OR "open-enrollment" OR "open enrollment")) OR (navigator* OR assistor* OR assister*) NEAR/10 ("health insurance" OR Medicare OR Medicaid) NEAR/5 enroll*)) OR ("CHIP" OR "Children’s Health Insurance Program") OR ("individual mandate" NEAR/10 "health insurance") OR "employer-sponsored insurance" OR (employer OR employee) NEAR/10 "health insurance")
Overall trial results

Categorized to a topic (Recall)

- Uncategorized: 33%
- Categorized: 67%

Categorized to the correct topic (Precision)

- Incorrect: 11%
- Correct: 89%
Precision and recall tradeoff
Trial results for each topic

**Childhood Obesity**
- Incorrect: 17%
- Correct: 83%
- n=64

**Disease Prevention and Health Promotion**
- Incorrect: 20%
- Correct: 80%
- n=65

**Health Care Quality**
- Incorrect: 4%
- Correct: 96%

**Health Coverage**
- Incorrect: 3%
- Correct: 97%
- n=71
Boolean search: How hard is it do?

**Machine learning**
- Machine learning is opaque.
  - Benefit: No preparation. Content just gets processed.
  - Problem: Categories are generic, may be irrelevant, can be biased, and are difficult to change or tune.

**Boolean queries**
- Pre-defined categories (aka a taxonomy) plus Boolean queries to scope the context for categories are transparent.
  - Benefit: Relevant categories.
  - Problem: Requires work to set up, and specialized skills.
Boolean queries

- Basic operators
  - AND (conjunctive)
  - OR (disjunctive)
  - NOT (negation)

- Venn diagrams
  - A OR B
  - A AND B
  - A NOT B
  - B NOT A
Proximity operators

- Proximity search (specified distance). **Hint:** Proximity operators and syntax are not standardized.
  - NEAR
  - NOT NEAR
  - FOLLOWED BY
  - NOT FOLLOWED BY
  - SENTENCE
  - FAR
Query syntax

❖ Bounded phrase
  ▪ Usually quotation marks, e.g.
    “health insurance”

❖ Truncation (right, left, internal)
  ▪ Usually an asterisk, e.g.
    child*
    “pre-existing condition*”

❖ Nested statements
  ▪ Parentheses (that must match up)
    (“health insurance” AND (children* OR “pre-existing condition*”))
How to create a Boolean query (1)

1) Brainstorm a list of 10 relevant words and phrases.
2) Use that list to identify 10 relevant items (articles, videos, websites, etc.)
   - E.g., do a Google search, search Google Scholar, search the NYT (or any other newspaper that you subscribe to), search Library of Congress Chronicling America (1789-1963), etc.
3) Review 10 relevant items and write down the words and phrases that provide a context for the theme/topic/concept.
   - Titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) are good areas to focus on without having to read the whole thing.
4) Note any named entities (people, organizations, events, laws, etc.) that are closely associated with the theme/topic/concept.
   - E.g., for gun violence Gabrielle Giffords, Michael Bloomberg, Doctors Against Gun Violence, March for our Lives, etc.
How to create a Boolean query (2)

5) Consolidate the terms.
   ▪ Identify duplicates, synonyms, as well as any concepts that you want to combine even if they are
     not synonyms.
   ▪ Re-label the term as needed to reflect the concept/category. Also consider and note any other
     relationships between terms. Prioritize the terms. Rank from 1-N, most relevant to least relevant.
     ▪ **Hint**: Rank each term by higher, medium, lower relevance, then sort and rank from 1-N.

6) Write a query for each term.
   ▪ Note that regular plurals (-s, -es, -ies) are usually (but not always) included automatically, but you
     always need to specify irregular plurals, e.g., “mice”.

7) Qualify the scope for each term.
   ▪ Does the term require any qualification of the scope, e.g., by population, setting, geography, etc.?
   ▪ Validate that the term is disjunctive, distinct, and requires no further qualification.

8) Combine the terms into a single nested query with an OR operator.
Activity: Create a Boolean search statement to scope one of the following concepts – Choose your level

<table>
<thead>
<tr>
<th>Your Level</th>
<th>Your Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>school shootings</td>
</tr>
<tr>
<td>Intermediate</td>
<td>public health</td>
</tr>
<tr>
<td>Advanced</td>
<td>what can public health do to curb school shootings</td>
</tr>
</tbody>
</table>
Questions, summary & evaluation

10 minutes

Joseph Busch,
jbusch@taxonomystrategies.com
joseph@semanticstaffing.com
m 415-377-7912
Appendix

- More information
- Cost-benefit analysis
- Pricing models
Cost benefit analysis: With and without automatic tagging

**With automation**
- **0 Minutes**: Save Content
- **1 Minute**: Automatic suggestion of categorization & metadata for content
- **2 Minutes**: User review and accepts or edits suggestions.
- **AUTOMATED**: Submit content for publishing

**Production Server**

**Cost of Labor = $100/hr**
**Total Time = 3:00 minutes**

**Cost per page = $5**

**Savings Per Page = $20**
**Pages = 100,000**

**Total Savings = $2,000,000**

**Without automation**
- **0 Minute**: Save content item
- **1 Minute**: Open metadata capture form
- **10 Minutes**: Apply metadata
- **3 Minutes**: Summarize content
- **1 Minute**: Save and submit content with metadata & summary

**Cost of Labor = $100/hr**
**Total Time = 15:00 minutes**

**Cost per page = $25**
Pricing example: Lexalytics
Pricing by documents processed

<table>
<thead>
<tr>
<th>Per month</th>
<th>Per year Full Price</th>
<th>Per Year Discounted</th>
<th>Transactions</th>
<th>Configurations</th>
<th>Excel Seats</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1,500</td>
<td>$18,000</td>
<td>$9,000</td>
<td>100,000</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>$2,500</td>
<td>$30,000</td>
<td>$15,000</td>
<td>1,000,000</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>$3,500</td>
<td>$42,000</td>
<td>$21,000</td>
<td>5,000,000</td>
<td>100</td>
<td>5</td>
</tr>
</tbody>
</table>

- Offers educational and non-profit pricing. "Pricing for educational institutions is 50% off the original package price. This can also be offered to certain non-profit organizations. [https://www.lexalytics.com/prices](https://www.lexalytics.com/prices).

- This model only considers the Semantria API topic queries method. Lexalytics offers other categorization methods including model based classifiers and concept matrix. We don't know what these other services might cost. [https://www.lexalytics.com/technology/categorization](https://www.lexalytics.com/technology/categorization)
Pricing example: IBM
Pricing by classification method calls - Natural Language Understanding

<table>
<thead>
<tr>
<th>Per month</th>
<th>Per year</th>
<th>Transactions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ 0.003</td>
<td></td>
<td>Per item</td>
<td>1-250,000 items</td>
</tr>
<tr>
<td>$ 0.001</td>
<td></td>
<td>Per item</td>
<td>250,001-5,000,000 items</td>
</tr>
<tr>
<td>$ 0.0002</td>
<td></td>
<td>Per item</td>
<td>5,000,000+ items</td>
</tr>
<tr>
<td>$ 150.00</td>
<td></td>
<td>Per user</td>
<td>Knowledge Studio</td>
</tr>
<tr>
<td>$ 800.00</td>
<td></td>
<td>Per custom model</td>
<td>Created in knowledge Studio</td>
</tr>
<tr>
<td>$ 150.00</td>
<td>$ 1,800.00</td>
<td></td>
<td>1 Knowledge Studio</td>
</tr>
<tr>
<td>$ 800.00</td>
<td>$ 9,600.00</td>
<td></td>
<td>1 Number of custom models</td>
</tr>
<tr>
<td>$ 120.00</td>
<td>$ 1,440.00</td>
<td></td>
<td>40000 Items per month</td>
</tr>
<tr>
<td></td>
<td>$ 12,840.00</td>
<td></td>
<td>Estimated annual cost</td>
</tr>
</tbody>
</table>

- This is a text analytics web service (comparable to Lexalytics Semantria)
- An item is 10,000 characters. Documents with greater than 10,000 characters are split into multiple 10K character items. Features include Categories, Concepts, Emotion, Entities, Keywords, Metadata, Relations, Semantic Roles, and Sentiment. Each feature in the API call is counted as a separate item.
- Only one custom model is required for RWJF Topics. These would be a specialization of the Categories feature. Documents are limited to 10,000 characters. Only 4 features are extracted per item. I.e., each document equals 4 items.
## Pricing example: IBM
### Pricing by classification method calls - Natural Language Classifier

<table>
<thead>
<tr>
<th>Unit cost</th>
<th>Per year</th>
<th>Unit</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ 20.00</td>
<td>$ 240</td>
<td>Per month</td>
<td>First classifier is free. Each concept equals 1 classifier.</td>
</tr>
<tr>
<td>$ 0.0035</td>
<td></td>
<td>Per call</td>
<td>1,000 free API calls per month. (See note)</td>
</tr>
<tr>
<td>$ 3.00</td>
<td></td>
<td>Per training event</td>
<td>First 4 training events per month are free.</td>
</tr>
<tr>
<td>$ 980.00</td>
<td>$ 11,760</td>
<td></td>
<td>Number of classifiers</td>
</tr>
<tr>
<td>$ 31.50</td>
<td>$ 378</td>
<td></td>
<td>API calls per month</td>
</tr>
<tr>
<td>$ 288.00</td>
<td>$ 288.00</td>
<td></td>
<td>Training events (assume one-time only)</td>
</tr>
</tbody>
</table>

|                  |      |                 | Estimated annual cost                                               |
|------------------|------|-----------------|
|                  | $    | 12,426          |

- This is a statistical categorizer.
- Assume 1 API equals 1 document, but could be per classifier per document, i.e., x 50.

More information: Overview


More information: Demos


More information: Challenges


- ProPublica. “Breaking the Black Box” series.