Supervised vs. Unsupervised Automated Categorization

Joseph Busch, Principal Analyst
Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study
Categorization goals

- Make sense
- Provide clear perception
- Is trustworthy and as unbiased as possible
Natural language processing enables unsupervised automated categorization
Individual terms can be replaced by pre-defined queries chosen and constructed by editors

- tf becomes qf
- Simple pre-defined query for “Affordable Care Act”
  - ("Affordable Care Act" OR "ACA" OR Obamacare)
- Complex pre-defined query for “Medicare Navigator”
  - ((navigator* OR assistor* OR assister*) NEAR ("health insurance" OR Medicare OR Medicaid) NEAR enroll*))
Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study
How reliable does a categorizer need to be?
It depends on the categorization target

Documents VS. Projects
Productivity vs risk models
Low vs high error tolerance

- High Value & Low Volume
- Low Value & Low Volume
- High Exposure & Low Risk
- Low Exposure & Low Risk
- High Value & High Volume
- Low Value & High Volume
- High Exposure & High Risk
- Low Exposure & High Risk
Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study
How to evaluate text analytics accuracy

- Build a prototype to test feasibility of auto-categorization.
  - Test selected categories individually
    - Is the recall and precision optimized on a category basis?
  - Test selected categories together
    - Can the categorizer appropriately distinguish between categories?
    - Are multiple categories identified when that is appropriate?
    - Are there relevance measures that can be meaningfully used to distinguish among multiple identified categories?

- Build production-quality categorizers.
  - Test all categories individually to optimize recall and precision
  - Test all categories together to ensure they can be appropriately distinguished
Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study
Former “A-Z” RWJF Taxonomy for their Project Information Management System (PIMS)

- Access and barriers to care
- Accountable care organizations
- Accreditation
- Addiction and substance abuse
- Administration
- Advanced practice nurses
- Affordable Care Act (ACA)
- Aging
- Alcoholism
- Allied health
- Alternative and spiritual care
- Ambulatory care
- American Indian tribal government
- Antibiotic resistance
- Asthma
- At-risk and vulnerable people
- Barriers to care: cultural, gender and racial
- Barriers to care: financial
- Barriers to care: language and literacy
- Barriers to care: logistics and transportation
- Behavior change
- Behavioral economics
- Behavioral/mental health
- Benchmarks and best practices
- Built environment
- Bundled payments
- Business
- Cancer
- Care transitions
- Careers
- Case management
- Child abuse and neglect
- Child care centers
- Child welfare
- Children's health insurance program
- Chronic disease management
- Chronic illness
- Clean air laws
- Clinical care
- Clinical research
- Colleges and universities
- Community benefit
- Community development
- Community outreach
- Community violence
- Community-based care
- Competitions
- Competitive foods
- Complete streets
- Conferences
- Consumer engagement
- Continuing education
- Coordinated care
- Cost of care
- Crowd sourcing
- Cultural competence
- Data
- Dental schools
- Dentists
- Developmental disabilities
- Diabetes
- Disabilities
- Disruptive innovations
- Diverse communities
- Diversity
- Early childhood
- Early intervention
- Education and training
- Education level
- E-health
- Eligibility and enrollment
- Emergency care
- Emergency preparedness and response
- Employers
- Employer-sponsored insurance
- End-of-life care
- Environmental health
- Epidemiology
- Epistomology
- Ethics/bioethics
- Evaluation
- Evidence-based
- Families
- Family caregiving
- Federal government
- Food access
- Food safety
- Foster care
- Frontline workforce
- Genetics
- Health care delivery system
- Health care markets
- Health education
- Health games
- Health impact assessments
- Health insurance
- Health insurance exchanges
- Health IT
- Health plans
- Health policy
- Health promotion and disease prevention
- Health records/electronic health records
- Health reform
- Healthy communities
- Home health care
- Home visiting
- Homeless
- Hospital readmissions
- Hospital-acquired infections
- Hospitals
- Housing/public housing
- Immigrants and refugees
- Immunizations
- Independent living or self-determination
- Individual insurance
- Industry regulation
- Infectious diseases
- Informatics
- Information technology
- Injury
- Inmates and former inmates
- Inpatient care
- Institutional design
- Insurance cost
- Interpersonal violence
- Interprofessional collaboration
- Jails and prisons
- Job satisfaction
- Jobs
- Joint use facilities
- Journalists
- Language services
- Leadership development
- Legal systems and issues
- Lesbian/gay/bisexual/transgender
- Local government
- Long-term care
- Low-birthweight infants
- Managed care organizations
- Marketing
- Medicaid
- Medical errors
- Medical homes
- Medical malpractice
- Medical practices
- Medical schools
- Medical students and residents
- Medical technology
- Medical, dental and nursing workforce
- Medically underserved areas
- Medicare
- Mentoring
- Military/veterans
- Mobile health/mhealth
- Models
- Networks
- Nurse executives
- Nurse midwives
- Nurse practitioners
- Nurses
- Nursing homes
- Nursing schools
- Nutrition
- Nutrition policy
- Obesity policy
- Open sourcing
- Oral health
- Out-of-school time
- Overweight
- Palliative care
- Parks or playgrounds
- Patient safety and outcomes
- Patient satisfaction
- Patient-centered care
- Patients
- Pay for performance
- Payment reform
- Pediatric care
- Performance standards and measurement
- Philanthropy
- Physical activity
- Physical activity policy
- Physical education
- Physician assistants
- Physician workforce
- Poor and economically disadvantaged
- Practice guidelines
- Prevention
- Preventive care
- Primary care
- Primary care/generalist physicians
- Primary care/preventive care
- Primary care/preventive care
- Privacy and confidentiality
- Provider incentives
- Public health
- Public health agencies
- Public health law
- Public health professionals
- Public health schools
- Public health system and finance
- Public policy
- Public-private partnerships
- Quality of care
- Recruitment and retention
- Registered nurses
- Religious congregations
- Research
- Research networks
- Risky behavior
- RWJF Transitions
- Safe routes to school
- Safety nets
- School and colleagues
- School foods
- School lunches
- School lunch programs
- School-based health centers
- Schools K-12
- Scope of practice
- Screening
- Self-care
- Shared decision making
- Shelters
- Shortage of medical or nursing personnel
- Small businesses
- Social and emotional learning
- Social determinants of health
- Social isolation
- Social marketing
- Social sciences
- Social support services
- Specialist physicians
- Specialist physicians
- State government
- Substance abuse treatment
- Substance use
- Substance use
- Sugary beverages
- Sugary beverages
- Supportive housing
- Tax policy
- Tax policy
- Telemedicine and telehealth
- Tobacco
- Tobacco
- Tobacco cessation
- Tobacco cessation
- Tobacco control
- Transdisciplinary
- Transparency/public reporting
- Transportation
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Transportation policy
- Underinsured
- Underinsured populations
- Underinsured populations
- Uninsured
- Value-based purchasing
- Walking and biking
- Walking and biking
- Walking and biking
- Work environment
- Workflow
- Workforce issues
- Workforce supply and demand
- Workforce supply and demand
- Youth development
New RWJF Taxonomy model for their Project Information Management System (PIMS)

<table>
<thead>
<tr>
<th>Topics</th>
<th>Activities</th>
<th>Geo Locations</th>
<th>Health Factors</th>
<th>Settings</th>
<th>Populations</th>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built Environment &amp; Health</td>
<td>Advocacy</td>
<td>+ Location Type</td>
<td>Access to Care</td>
<td>American Indian</td>
<td>Activist/Advocate</td>
<td>+ Age</td>
</tr>
<tr>
<td>Child &amp; Family Well-being</td>
<td>Capacity &amp; Organizational Development</td>
<td>+ Regions</td>
<td>Air &amp; Water Quality</td>
<td>Tribal Government</td>
<td>Administrative/Managerial</td>
<td>+ Gender</td>
</tr>
<tr>
<td>Childhood Obesity</td>
<td>Coalition &amp; Network Building</td>
<td>+ States &amp; Territories</td>
<td>Alcohol &amp; Drug Use</td>
<td>Business (General)</td>
<td>Caregiver</td>
<td>+ Education</td>
</tr>
<tr>
<td>Disease Prevention &amp; Health</td>
<td>Communications &amp; Media</td>
<td></td>
<td>Community Safety</td>
<td>Child Care</td>
<td>Faith-based</td>
<td>+ Income Level</td>
</tr>
<tr>
<td>Promotion</td>
<td>Evaluation</td>
<td>Diet &amp; Exercise</td>
<td>Community (General)</td>
<td>Education</td>
<td>Health Professional</td>
<td>+ Race &amp; Ethnicity</td>
</tr>
<tr>
<td>Early Childhood</td>
<td>Idea Sourcing</td>
<td>Employment &amp; Income</td>
<td>Federal Government</td>
<td>Family &amp; Social Support</td>
<td>Immigrant or Migrant</td>
<td></td>
</tr>
<tr>
<td>Health Care</td>
<td>Leadership, Education &amp; Training</td>
<td>Genetics</td>
<td>Health Care/Clinical</td>
<td>Support</td>
<td>Inmates or Former Inmates</td>
<td></td>
</tr>
<tr>
<td>Coverage &amp; Access</td>
<td>Pilot Project</td>
<td>Housing, Transit &amp; Infrastructure</td>
<td>Home/Household</td>
<td>Health</td>
<td>LGBTQ Person</td>
<td></td>
</tr>
<tr>
<td>Health Care Quality &amp; Value</td>
<td>Research &amp; Analysis</td>
<td>Quality of Care</td>
<td>Local Government</td>
<td>Disparities</td>
<td>Non-Health Professional</td>
<td></td>
</tr>
<tr>
<td>Health Disparities</td>
<td></td>
<td>Tobacco Use</td>
<td>Nonprofit/Philanthropy</td>
<td>Leadership, Education &amp; Training</td>
<td>Parent</td>
<td></td>
</tr>
<tr>
<td>Health Leadership Development</td>
<td></td>
<td></td>
<td>School (General)</td>
<td>Research &amp; Analysis</td>
<td>Person with Disability</td>
<td></td>
</tr>
<tr>
<td>Nurses and Nursing</td>
<td></td>
<td></td>
<td>State Government</td>
<td></td>
<td>Person with Disease or Illness</td>
<td></td>
</tr>
<tr>
<td>Public &amp; Community Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Policymaker</td>
<td></td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student</td>
<td></td>
</tr>
</tbody>
</table>

**Problem:** How to convert legacy collection from as-is to the new taxonomy model?
- Option 1: Mapping from the as-is to the new.
- Option 2: Automated categorization
- Option 3: Combination of option 1 and 2
## Option 1: Mapping

<table>
<thead>
<tr>
<th>CODING _DEF_ID</th>
<th>PIMS Topics</th>
<th>New Facet</th>
<th>New Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>Access and barriers to care</td>
<td>Topics</td>
<td>Health Care Coverage &amp; Access</td>
</tr>
<tr>
<td>1707</td>
<td>Accountable care organizations</td>
<td>Topics</td>
<td>Health Care Quality &amp; Value</td>
</tr>
<tr>
<td>386</td>
<td>Accreditation</td>
<td>Topics</td>
<td>Public &amp; Community Health</td>
</tr>
<tr>
<td>508</td>
<td>Addiction and substance abuse</td>
<td>Health Fac Alcohol &amp; Drug Use</td>
<td></td>
</tr>
<tr>
<td>670</td>
<td>Administration</td>
<td>Population Administrative/Managerial</td>
<td></td>
</tr>
<tr>
<td>670</td>
<td>Administration</td>
<td>Population Health Professional</td>
<td></td>
</tr>
<tr>
<td>670</td>
<td>Administration</td>
<td>Topics</td>
<td>Health Leadership Development</td>
</tr>
<tr>
<td>470</td>
<td>Advanced practice nurses</td>
<td>Population Health Professional</td>
<td></td>
</tr>
<tr>
<td>470</td>
<td>Advanced practice nurses</td>
<td>Topics</td>
<td>Nurses &amp; Nursing</td>
</tr>
<tr>
<td>1709</td>
<td>Affordable Care Act (ACA)</td>
<td>Topics</td>
<td>Health Care Coverage &amp; Access</td>
</tr>
<tr>
<td>1667</td>
<td>Aging</td>
<td>Topics</td>
<td>Demographics Older Adults</td>
</tr>
<tr>
<td>509</td>
<td>Alcoholism</td>
<td>Health Fac Alcohol &amp; Drug Use</td>
<td></td>
</tr>
<tr>
<td>509</td>
<td>Alcoholism</td>
<td>Population Person with Disease or Illness</td>
<td></td>
</tr>
<tr>
<td>509</td>
<td>Alcoholism</td>
<td>Topics</td>
<td>Disease Prevention &amp; Health Promotion</td>
</tr>
<tr>
<td>478</td>
<td>Allied health</td>
<td>Population Health Professional</td>
<td></td>
</tr>
<tr>
<td>478</td>
<td>Allied health</td>
<td>Topics</td>
<td>Health Leadership Development</td>
</tr>
<tr>
<td>406</td>
<td>Alternative and spiritual care</td>
<td>Settings</td>
<td>Health Care/Clinical</td>
</tr>
<tr>
<td>395</td>
<td>Ambulatory care</td>
<td>Settings</td>
<td>Health Care/Clinical</td>
</tr>
<tr>
<td>395</td>
<td>Ambulatory care</td>
<td>Topics</td>
<td>Health Care Quality &amp; Value</td>
</tr>
<tr>
<td>306</td>
<td>American Indian tribal government</td>
<td>Settings</td>
<td>American Indian Tribal Government</td>
</tr>
</tbody>
</table>
Option 2-Step 1: Build a prototype to test feasibility of auto-categorization
Can we build some categorizers that accurately detect Topics?

- Collected 200 categorized examples for each Topic to be tested.
- Divided test collection in half – 100 to use to build the categorizer, and 100 to use in a blind test.
- Reviewed 100 relevant items and identify the words and phrases that provide a context for the topic.
- Noted any named entities (people, organizations, events, laws, etc.) that are closely associated with the topic.
- Consolidated the terms – identify duplicates, synonyms, as well as any concepts that you want to combine even if they are not synonyms.
- Wrote a query for each term.
- Combined the terms into a single nested query.
Pre-defined query for “Health Care Coverage & Access” topic

("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 (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")
Trial results for each pre-defined category

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

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

**Health Care Quality & Value**
- Correct: 96%
- Incorrect: 4%
- n=69

**Health Care Coverage & Access**
- Correct: 97%
- Incorrect: 3%
- n=71
Option 2-Step 2: Build production-quality categorizers for all 12 topics

- For each of the 12 RWJF Topics:
  - Gather a test collection of grant summaries
  - Build and test a rule – a Boolean query with proximity operators to classify into an RWJF Topic
    - Start with the RWJF Topic string.
    - Add synonyms.
    - Add legacy PIMS Topics that were mapped to the 12 RWJF Topics and synonyms.
    - Review the test collection of grant titles & summaries tagged for that RWJF Topic. (A sample size of 100 should suffice). Identify distinctive words and phrases.
    - Group Boolean search strings together by each discrete topic as much as possible.
    - Review and revise.
  - Test a categorizer for an RWJF Topic. (12 categorizers; 12 tests)
    - Test recall – Are all the relevant grants being identified by the query, 70-80% of the time? If not, revise query and re-test.
    - Test precision – Are only the relevant grants being identified by the query, 70-80% of the time? If not, revise query and re-test.

- Consolidate 12 RWJF Topics into a master categorizer.
  - Test master categorizer against NEW test collection.
Health Care Coverage > Health Insurance Characteristics sub-queries
(Note: we tested at the whole Topic level, not the sub-queries)

- **Health Care Access**
  - Access to Health Care
  - Barriers to Health Care
  - Various Health Outcomes Based on Access

- **Health Care Coverage**
  - Health Care Reform – Legislation
  - Health Insurance Characteristics
    - Health Insurance Cost
    - High Deductible Insurance
    - High Risk
    - Mental Health-Substance Abuse
    - PPO-HMO
    - Pre-existing Condition
    - Provider Network-In-Out
  - Health Insurance Enrollment
  - Insurance Marketplace

  - ("health care" OR healthcare OR "health insurance") AND (premium* OR payment* OR rate* OR cost*)
  - (high-deductible OR "high deductible") NEAR/5 "health insurance"
  - (high-risk OR "high risk") NEAR/5 "health insurance"
  - ("mental health" OR "substance abuse") NEAR/5 "health insurance"
  - (PPO* OR HMO*) NEAR/5 (marketplace* OR plan* OR provider*)
  - ("preexisting condition**" OR "pre-existing condition**") NEAR/5 "health insurance"
  - ("provider network**" NEAR/5 "health insurance") OR (("in-network" OR "out-of-network") NEAR/5 "health insurance"))
Optimize recall and precision on a category-by-category basis

<table>
<thead>
<tr>
<th>Topic</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built Environment &amp; Health</td>
<td>78%</td>
<td>70%</td>
</tr>
<tr>
<td>Child &amp; Family Well-being</td>
<td>77%</td>
<td>74%</td>
</tr>
<tr>
<td>Childhood Obesity</td>
<td>86%</td>
<td>97%</td>
</tr>
<tr>
<td>Disease Prevention &amp; Health Promotion</td>
<td>57%</td>
<td>64%</td>
</tr>
<tr>
<td>Early Childhood Development</td>
<td>76%</td>
<td>82%</td>
</tr>
<tr>
<td>Health Care Coverage &amp; Access</td>
<td>72%</td>
<td>79%</td>
</tr>
<tr>
<td>Health Care Quality &amp; Value</td>
<td>69%</td>
<td>75%</td>
</tr>
<tr>
<td>Health Disparities</td>
<td>73%</td>
<td>81%</td>
</tr>
<tr>
<td>Health Leadership Development</td>
<td>71%</td>
<td>80%</td>
</tr>
<tr>
<td>Nurses and Nursing</td>
<td>94%</td>
<td>94%</td>
</tr>
<tr>
<td>Public &amp; Community Health</td>
<td>51%</td>
<td>69%</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>71%</td>
<td>69%</td>
</tr>
</tbody>
</table>
Option 2-Step 3: Consolidate category-by-category queries into a master categorizer

- Combine all of the category queries into one master categorizer
  - Make changes to queries as needed to meet the parameters of a single categorizer
  - Add named entities where appropriate

- Run analysis and calculate precision and recall to discover if any unintended problems were introduced due to consolidating into one categorizer
  - Use same test set of assets (Step 2 test set) for Step 3
Optimize recall and precision for consolidated categorizer

... is this performance reliable enough for unsupervised categorization of high risk content?
Option 3: Combination of mapping and automated methods

- Used mapping to provide initial conversion of legacy content from A-Z to new faceted taxonomy.
- Identified 1) categories with no mapping to any grants, and 2) grants with too many Topics; then processed these grants with a combination of human review, and automated categorization to improve their indexing.
- Worked to build a critical mass of new grants indexed by editors using the new taxonomy.
- **Future goal:** Automation-assisted indexing.
Summary

- Natural language processing enables unsupervised automated categorization.
- Individual terms can be replaced by pre-defined queries chosen and constructed by editors.
- Good results on individual pre-defined queries/topics may not be as good when distinguishing which is most relevant among a group of pre-defined queries/topics.
- Automated categorization results may be good enough for some collections of content (high volume or low risk), but not for others (low value or high risk).
Questions

Joseph Busch
jbusch@taxonomystrategies.com or joseph@semanticstaffing.com
@joebusch
mobile +1 415-377-7912
Supervised vs. Unsupervised Automated Categorization: Summary

- AI promises to categorize all types of content with reliable results, but the reality is much more complex. Most applications won’t work with a meat grinder approach where you pour a huge amount of content in one end and a perfectly organized collection comes out the other end. Effective automated categorization depends on defining a process workflow and assembling a stack of methods to process different types of content in different ways. Designing and validating a content processing workflow requires human judgements. So good quality categorization applications often rely on how to make the best use of people. This presentation provides a reality check on unsupervised automated categorization, and discusses a case study where the performance was suitable for editorial review and approval, but not for unsupervised processing of a large collection. The project is the Robert Wood John Foundation grants information system using Lexalytics configurations developed by Taxonomy Strategies.