Computer-Assisted Text Analysis: An Overview and Guide

Laura K. Nelson
Kellogg School of Management
Content Analysis PWD
Academy of Management Annual Conference
August 7, 2015
Vancouver, BC
Goals

- Describe the wide range of computer-assisted text analysis techniques available
  - Hint: more than dictionaries
- Provide some guidance about how to choose your methods
- Give empirical examples of these methods in practice
Why Use Computer-Assisted Techniques?

- Speed
  - Humans are slow
  - Text is becoming large
- Reliability / Reproducibility
- Validity (this is controversial)
  - Expanded memory
  - Unburdened by bias

Does not remove the need for interpretation!
Overview:
Types of Automated Text Analysis

- Unsupervised exploration (hypothesis forming/inductive)
  - Topic modeling
  - Lexical selection
- Human-Guided Categorical Analysis (traditional content analysis – deductive hypothesis testing)
  - Supervised machine learning
  - Dictionaries
- Natural Language Processing (guided inductive/hypothesis testing)
  - Part-of-Speech Tagging
  - Named Entity Recognition
  - Concordances
  - Sentiment analysis
Question 1:
Do you want to inductively explore the text?
Does each document contain multiple themes?

- Yes: Do you have meta-data that meaningfully connects document?
  - Yes: STM
  - No: LDA

- No: Clustering Algorithm (e.g. K-Means)

Do you want to explore themes, or the way themes appear?

- Yes: TF-IDF Scores
  - Yes: The way in which themes are discussed
    - Yes: Do you have more than two categories?
      - Yes: Difference of Proportions
      - No: Difference of Proportions
    - No: Difference of Proportions

- No: Word Scores

The way in which themes are discussed

- Yes: Do you want to do this inductively?
Unsupervised Exploration: The Goal

Informative Groups of Words
### Set-Up: Document-Term Matrix*

<table>
<thead>
<tr>
<th></th>
<th>ambit</th>
<th>poverti</th>
<th>people</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document2</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Document3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Document4</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Document5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

*Cells can be word frequencies or weighted word scores*
Question 2: Themes or Style?

If themes:

Question 3: Multiple Categories?
Single Category per Text: Clustering
Multiple Categories: Topic Modeling

Seeking Life's Bare (Genetic) Necessities

Cold Spring Harbor, New York—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic gene needs of life.

One research team, using computer models, says that using less than 1000 genes is enough. The other researchers, mapping genes in a simple parasite, estimated that they need at least 3000 genes, and that the earliest life forms required a mere 125 genes. The other researchers mapped genes in a simple parasite and estimated that 300 genes are enough for the job—"but that anything short of 100 would be enough." Although the numbers don't match precisely, those findings are consistent with the idea that "not all that far apart," especially in comparison to the 3000 genes in the human genome and the 300 in the yeast genome.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.

"It's not a question of how many genes but how diverse they are," says Andrew D. Smith, a biology professor at the University of Colorado. "There may be more than one way to make a living." The researchers are using their genes to make a living, and the genes are being sequenced. "It may be a way of organisms and other organisms, particularly the ones that are more ancient, to make a living," says Andrew D. Smith.
But Which Algorithm?

- Is the order of the documents important?
  - Yes? Structural Topic Modeling (STM)
- Are the topics correlated?
  - Yes? Correlated Topic Modeling (CTM)
- Order is relatively arbitrary, topics may not be related?
  - Latent Dirichlet Allocation (LDA)
<table>
<thead>
<tr>
<th>music</th>
<th>band</th>
<th>songs</th>
<th>rock</th>
<th>album</th>
<th>jazz</th>
<th>pop</th>
<th>song</th>
<th>singer</th>
<th>night</th>
</tr>
</thead>
<tbody>
<tr>
<td>book</td>
<td>life</td>
<td>novel</td>
<td>story</td>
<td>books</td>
<td>man</td>
<td>stories</td>
<td>love</td>
<td>children</td>
<td>family</td>
</tr>
<tr>
<td>art</td>
<td>museum</td>
<td>show</td>
<td>exhibition</td>
<td>artist</td>
<td>artists</td>
<td>paintings</td>
<td>painting</td>
<td>century</td>
<td>works</td>
</tr>
<tr>
<td>game</td>
<td>knicks</td>
<td>nets</td>
<td>points</td>
<td>team</td>
<td>season</td>
<td>play</td>
<td>games</td>
<td>night</td>
<td>coach</td>
</tr>
<tr>
<td>show</td>
<td>film</td>
<td>television</td>
<td>movie</td>
<td>series</td>
<td>says</td>
<td>life</td>
<td>man</td>
<td>character</td>
<td>know</td>
</tr>
<tr>
<td>theater</td>
<td>play</td>
<td>production</td>
<td>show</td>
<td>stage</td>
<td>street</td>
<td>broadway</td>
<td>director</td>
<td>musical</td>
<td>directed</td>
</tr>
<tr>
<td>clinton</td>
<td>bush</td>
<td>campaign</td>
<td>gore</td>
<td>political</td>
<td>republican</td>
<td>dole</td>
<td>presidential</td>
<td>senator</td>
<td>house</td>
</tr>
<tr>
<td>stock</td>
<td>market</td>
<td>percent</td>
<td>fund</td>
<td>investors</td>
<td>funds</td>
<td>companies</td>
<td>stocks</td>
<td>investment</td>
<td>trading</td>
</tr>
<tr>
<td>restaurant</td>
<td>sauce</td>
<td>menu</td>
<td>food</td>
<td>dishes</td>
<td>street</td>
<td>dining</td>
<td>dinner</td>
<td>chicken</td>
<td>served</td>
</tr>
<tr>
<td>budget</td>
<td>tax</td>
<td>governor</td>
<td>county</td>
<td>mayor</td>
<td>billion</td>
<td>taxes</td>
<td>plan</td>
<td>legislature</td>
<td>fiscal</td>
</tr>
</tbody>
</table>
Question 2: Themes or Style?

If style...
Style: Lexical Selection

- Goal: find words that are distinctive to different groups of text
- One solution: Difference of Proportions
Difference of Proportions

Chicago
- children: 4.59
- center: 4.34
- union: 3.61
- school: 3.48
- abort: 3.19
- nixon: 2.93
- day: 2.86
- vietnam: 2.57
- people: 2.50
- city: 2.44
- hospital: 2.38
- cwlu: 2.37

New York City
- movement: 12.54
- women: 11.34
- feminist: 8.91
- radical: 8.56
- liberation: 7.69
- political: 5.81
- history: 5.68
- feminine: 3.85
- male: 3.52
- left: 2.96
- revolution: 2.58
- consciousnessraising: 2.45
- oppress: 2.41

Abstract

Concrete
Question 1:
Do you want to test a hypothesis?

If yes:

Question 2: Themes or Styles?

If themes...
Supervised Machine Learning
Which Algorithm?

• You want individual documents coded:
  • Document Classification (e.g. SVM, Naive Bayes)

• You want proportion of documents in each category:
  • ReadMe (R package)
Dictionary Methods

- Standardized Dictionaries
- LIWC (can be used for sentiment analysis)
- Custom Dictionary
Question 2: Themes or Styles?

If styles...
is the frequency of the term important?

Yes → Concordances

No → Is the way the term is used important?

Yes → Part of Speech Tagger

No → Is grammar important?

Yes → Named Entity Recognition

No → Are the attributes of the words important?

Yes → Are Named Entities important?

Yes → WordNet

No → What about Synonyms? Antonyms? Hypernyms?

Yes → Other NLP techniques

No → Other NLP techniques

No → NLP for Guided Exploration & Hypothesis Testing
Natural Language Processing

- Takes into account features of words, relationships between words, grammatical structures, etc.
Examples: Use NLP to test hypotheses

• Hypothesis: Author A is more descriptive than Author B.
  • Test: Part-of-Speech tagger, extract adjectives, count and compare.

• Hypothesis: Organizations in New York City are more internationally focused than organizations in Silicon Valley.
  • Test: Named Entity Recognition, compare against lists of corporations, places, and people.
Example: Use NLP to test hypotheses

- Hypothesis: the word “disruptive” is used in a positive way, and has a different meaning, for Silicon Valley organizations compared to Wall Street organizations.
- Test: Concordances
NLP: Concordances

very heartily so exceedingly remarkably as vast a great amazingly
extremely good sweet
• Mostly positive

mean part maddens doleful gamesome subtly uncommon careful
untoward exasperate loving passing mouldy christian few true
mystifying imperial modifies contemptible
• Mostly negative
Example: NLP and WordNet

- Hypothesis: Women’s movement organizations in New York City approach politics more abstractly compared to those in Chicago, who have a more concrete approach to politics.
Tactics and Issues Over Time

• Structural Topic Models (structured on year)
  • Used R package *stm* (Roberts, Stewart, and Tingly)
• Further grouped the 40 topics into 7 topic categories
• Python NLTK, extracted verbs/verb phrases
• Hand identified tactics, created dictionaries of tactical categories
Tactics by Year
SUBARU AND THE ENVIRONMENT:
a Partnership FOR A Healthier Planet
Conclusion:

- Research design is key! Good data is critical!
- Match your method to your question and data. Be purposeful, not trendy
- Use multiple methods, including qualitative, to verify the analysis
- Learn a programming language
  - Off-the-shelf tools box you in (see point 2).
  - I recommend Python, R is also good
- Read NLP and machine learning literature
Happy text analyzing!

Laura K. Nelson
laura.nelson@kellogg.northwestern.edu
Tactical Categories*

Direct Environmental Protection: build, improve, protect, recycle
Non-Disruptive Protest: chant, demonstrate, organize, petition, protest
Disruptive Protest: blockade, chain, prevent, damage, sabotage
Political: campaign, donate, elect, endorse, regulate
Juridical: audit, enforce, inspect, represent, testify
Verbal Statements: advocate, comment, criticize, explain, refute
Business: boycott, buy, invest, purchase, sponsor
Education/Raising Awareness: editorial, outreach, publish, report, tweet
Organization/Movement Building: fund-raise, initiate, launch, participate
Negotiations: deal, discuss, engage, listen, persuade

*Categories are not mutually exclusive