TEXT MINING CHALLENGES AND SOLUTIONS IN BIG DATA

Dr. Normand Péladeau
President & CEO
Provalis Research Corp.
peladeau@provalisresearch.com

Dr. Derrick L. Cogburn
HCSS Global Virtual Teams Mini-Track Co-Chair
HCSS Text Analytics Mini-Track Co-Chair
Associate Professor, School of International Service
Executive Director, Institute on Disability and Public Policy
COTELO: The Collaboration Laboratory
American University
dcogburn@american.edu
@derrickcogburn

Recommended Texts
• An Introduction to Data Science, Jeffrey Stanton (2013). [Free iBook or PDF]
http://jsresearch.net/wiki/projects/teachdatascience
• Text Mining for Qualitative Data Analysis in the Social Sciences. Gregor Wiedemann (2016). Springer.

Understanding the “big” in Big Data

- Kilobyte (KB) = 1,024 bytes (2.3 paragraphs of plaintext)
- Megabyte (MB) = 1,048,576 bytes or 1,024 Kilobytes (873 pages of plaintext)
- Gigabyte (GB) = 1,073,741,824 bytes, 1,048 Megabytes, or 1,048,576 Kilobytes (894,784 pages of plaintext)
- Terabyte (TB) = 1,099,511,627,766 bytes, 1,024 Gigabytes, or 1,048,576 Megabytes (916,729,669 pages of plaintext)
- Petabyte (PB) = 1,125,899,906,842,624 bytes, 1,024 Terabytes, 1,048,576 Gigabytes, or 1,073,741,824 Megabytes (938,249,922,368 pages of plaintext)
- Exabyte (EB) = 1,152,921,504,606,846,976 bytes, 1,024 Petabytes, 1,048,576 Terabytes, 1,073,741,824 Gigabytes, or 1,099,511,627,766 Megabytes (960,767,920,505,705 pages of plaintext)
- Zettabyte (ZB) = 1,180,391,202,130,342,792 bytes, 1,024 Exabytes, 1,048,576 Petabytes, 1,073,741,824 Terabytes, 1,099,511,627,766 Gigabytes, or 1,125,899,906,842,624 Megabytes (983,826,350,597,842,752 pages of plaintext)
- Yottabyte (YB) = 1,208,925,189,614,628,174,706,776 bytes, 1,024 Zettabytes, 1,048,576 Exabytes, 1,073,741,824 Petabytes, 1,099,511,627,766 Terabytes, 1,125,899,906,842,624 Gigabytes, or 1,152,921,504,606,846,976 Megabytes (1,007,488,153,212,110,796,031 pages of plaintext)

Source: http://www.computerhope.com/vocab/chkpsz.htm

Defining the “big” in Big Data

- “Big Data” is a relative term
  - it means different things to different people/disciplines:
    • When we talk of “Big” data, we mean “big” less in absolute terms and more in terms relative to the comprehensive nature of the data.
    • 75-80% of the world’s available data is unstructured text (unstructured information growing at 15 times structured)
    • “In the past 50 years, the New York Times produced 3 billion words” and “Twitter users produce 8 billion words – every single day” (Kalev Leetaru, University of Illinois, and Kaisler, Armour, Espinosa, and Money, 2014)
    • In addition to text (websites, blogs, social media, email archives, annual reports, meeting transcripts, published articles – newspapers and journals) there are image, video, audio, GPS, RFID, and other types of Big Data

Understanding the “data” in Big Data

- Bits = Binary Digit (0, 1)
- Nibble = 4 bits
- Byte = 8 bits (256 combinations)
  – (NB: KB v. KB storage v. transmission)

<table>
<thead>
<tr>
<th>Meaning</th>
<th>2nd Digit</th>
<th>1st Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maybe</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Probably</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Definitely</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Source: Jeff Stanton, Introduction to Data Science

Objectives

At the end of this workshop, participants should be able to:

1. Understand the main challenges text analysts are facing.
2. Identify various text analysis strategies and techniques to deal with those challenges.
3. Recognize their respective strengths and weaknesses.
4. Identify various exploratory text mining techniques.
5. Apply dictionary construction and validation principles.

And if enough time
6. Understand some of the basic features of automatic document classification techniques.
**Defining the “big” in Big Data**

- The “Three Vs Model” of Big Data
  - Source: Doug Laney, Business Analyst, Gartner
- **Volume** = the amount of available data
- **Velocity** = speed at which data arrives/decays
- **Variety** = different types of data
  - Plus:
- **Veracity** = accuracy of the data
- **Variability** = differing interpretations of the data
- **Value** = relative importance of the data

**Acquiring Text Data**

- Tools and Techniques for scraping sites
  - Software
    - Sitesucker*
    - PageSucker
    - WebGrabber
    - Web Dumper
  - Hand Coding
    - Perl
    - Python
    - Ruby
- Downloading email listserv archives
  - Mbox format, Gmail, etc.
- Twitter APIs
  - dev.twitter.com
- Provalis Social Media Scraper
  - WebCollector
- Newspapers and published articles
  - e-library resources, Lexis-Nexis, etc.

**Text Analytics Applications**

- Sentiment Analysis (social media)
- Voice of the Customer (emails, chat, call center transcripts)
- Product improvement (warranty claims)
- Competitive Intelligence (patents, web sites)
- Risk management (incident or maintenance reports)
- Fraud detection (insurance claims)
- Reputation management (news, blogs, social media)
- Scientometrics studies (journal articles, titles & abstracts)
- Crime analysis (narratives, computer forensics, testimonies)
- Survey analysis (open-ended questions)
- Financial prediction (earnings releases, news, press releases)
- Surveillance system (communication, medical reports)
- Many more...

**Text Mining in the World Data Sciences**

**Text Analysis Landscape**

FOUR APPROACHES TO TEXT ANALYSIS

1. Computer Assisted Qualitative Analysis
2. Exploratory text mining
3. Quantitative Content Analysis
4. Automatic Document Classification

**Our Proposal**

- Each text analysis method has its own strengths and weaknesses.
- No single method is appropriate for all text analysis tasks.
- A single text analysis task often benefit from combining several methods.
### Tools we will use

<table>
<thead>
<tr>
<th>Content Analysis &amp; Text Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools</strong></td>
</tr>
<tr>
<td>WordNet</td>
</tr>
<tr>
<td>Qualitative Analysis &amp; Mixed Methods</td>
</tr>
</tbody>
</table>

### Some other tools available

**Commercial tools**
- IBM Text Modeler for Text, IBM Watson
- SAS Text Miner, Clarabridge, Lexalytics, AlchemyAPI, Attensity, Enkata, OdinText, etc.

**Open-source tools**
- Text mining modules R programming modules (R/tm), Gensim, Mallet, Quanteda, Rapid Miner, Gate, KNIME

**NLP Libraries**
- Stanford NLP, Natural Language Toolkit (NLTK), OpenCalais, Apache OpenNLP, etc.

http://www.kdnuggets.com/software/text.html

### Computer Assisted Qualitative Analysis

<table>
<thead>
<tr>
<th>Necessity is the mother of invention</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Laziness</strong></td>
</tr>
</tbody>
</table>

### Computer Assisted Qualitative Analysis

**Text Analytics Challenge**

**THREE MAJOR OBSTACLES**

1) Very large number of word forms

2) Polymorphy of language
   - One idea → multiple forms

3) Polysemy of words
   - One word → many ideas
Three Major Obstacles

1) Very large number of word forms
   - Polymorphy of language
     - One idea → multiple forms
2) Polysemy of words
   - One word → many ideas

Text Analytics Challenge

Challenge #1 – Quantity

38,996 comments about hotels
- 2,1 million words (tokens)
- 20,116 terms or word forms (types)

1,8 million course evaluations
- 35 millions words (tokens)
- 78,159 terms or word forms (types)

The “bag of words” assumption

- The order of the words in the document does not matter
- While a “big assumption” text mining experts have found that they can still differentiate between semantic concepts by using all the words in the documents
- Do not work in all situations and some information extraction tasks and natural language processing relies heavily on the words themselves (e.g. part of speech tagging) and the order of the words (preceding and following)
- Specialized algorithms are used in these cases
### Challenge #1 – Solutions

**Linguistic Pre-processing**
- Removal of stop words
- Stemming
- Lemmatization

**Statistical tools**
- Frequency selection
- Data reduction techniques (HCA, PCA, FA)
- Exploratory data analysis (ex. CA).
- Machine Learning

### Challenge #1 – Stop Words

**Removal of Stop Words**
- Words that are either insignificant (i.e., articles, prepositions) or too common
- Examples: “the”, “and”, “or”, “a”, “of”, “to”, “at”, “is”, “it”, “have”, “who”, etc.
- Caution: use with care
  - “IT” as “information technology”
  - “The Who”, “take that”, “a must”
  - Negation: not, no, never, seldom, no, etc
  - “but”, “however”, “otherwise”

### Challenge #1 – Stemming

Stemming - Removal of common prefixes and suffixes to obtain a word stem
- Example: prefix – stem – suffix
  - un – avail – able
- Issue: Stemming errors
  - universal, university, universe -> univers
  - designate, design -> design
  - paste, past -> past
  - political, polite -> polit
  - several, severance -> sever

### Challenge #1 – Lemmatization

Lemmatization: Reducing inflected forms of words to their canonical form.
- Examples: walk, walks, walked, waking -> walk
  - am, are, is -> be
- Two forms:
  1. Linguistic (very slow but more precise)
  2. Statistical (fast but less accurate)
- Issue - Some loss in semantic precision
  - Different uses of singular vs plural forms
  - Different uses of verb tenses

### The Statistics of Text

**Distribution of words: Zipf distribution**

![Zipf distribution graph]
The Statistics of Text

38,988 comments about hotels
2.1 M words (20,114 different terms)

<table>
<thead>
<tr>
<th>MOST FREQUENT TERMS</th>
<th>PERCENTAGE OF TERMS</th>
<th>PERCENTAGE OF WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>0.24%</td>
<td>50%</td>
</tr>
<tr>
<td>300</td>
<td>1.5%</td>
<td>76%</td>
</tr>
<tr>
<td>500</td>
<td>2.5%</td>
<td>83%</td>
</tr>
<tr>
<td>1000</td>
<td>5.0%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Text Analytics Challenge

TFxIDF – Term frequency x inverse document frequency

Heuristic technique for selecting words that are important in a corpus

Principles:
If a word appears frequently in a document, it’s important
If a word appears in many documents, it is less important

Basic formula: \( f_{t,d} \times \log \left( \frac{N}{n_t} \right) \)

Hierarchical Clustering

PROS
- Identification of topics & structure of topics
- May be used to reduce dimensionality
- Tends to group synonyms (polymorphy)

CONS
- Does not deal adequately with polysemy of words
- No single best solution
  (more later)

Topic Modeling (LSA, pLSA, LDA, PAM, etc.)

PROS
- Fast identification of topics
- Reduce dimensionality
- Deal partially with synonymy & polysemy

CONS
- No single best solution
  (more later)

Text Mining Approach

PROS
- Very fast
- Very little efforts
- Inductive

CONS
- Comparability of results
- Imprecise quantification
- Insensitive to low frequency events
- Sensitive to structured text elements
- Inductive

Text Analytics Challenge

THREE MAJOR OBSTACLES

1) Very large number of word forms
THREE MAJOR OBSTACLES

1) Very large number of word forms

2) Polymorphy of language
   One idea → multiple forms

SDG Goal 4: Education = 2

- 4.5 By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations

  - 4.a Build and upgrade education facilities that are child, disability and gender sensitive and provide safe, nonviolent, inclusive and effective learning environments for all

Goal 4.a
- Facilities
- Safe Environment
- Inclusive Environment
- Effective Learning
Creation of a dictionary

Custom Dictionary for SDGs and PWDs

Content analysis method

PROS

• Can potentially measure more accurately
• Can be focused (multi-focus)
• Allows full automation
• Allows comparison (overtime – across text collections)
• Allows measurements of latent dimensions
• Publicly or commercially available dictionaries
• Deductive approach

Measure Latent Dimensions

PSYCHOMETRIC MEASUREMENT

• Linguistic Inquiry and Word Count (LIWC) - Pennebaker
• Regressive Imagery Dictionary (RID) – Martindale
• Communication Vagueness Dictionary – Hiller
• Brand Personality Dictionary - Opoku

SOCIO-POLITICAL MEASUREMENT

• DICTON - Hart
• Lasswell Value Dictionary - Lasswell
• General Inquirer Harvard IV - Stone

COMMUNICATION VAGUENESS DICTIONARY

Measure Latent Dimensions
### Content analysis method

**PROS**
- Can potentially measure more accurately
- Can be focused (multi-focus)
- Allows full automation
- Allows comparison (overtime – across text collection)
- Allows measurements of latent dimensions
- Publicly or commercially available dictionaries
- Deductive

**CONS**
- Time required for construction & validation
- Improper use of existing dictionaries

---

### Text Analytics Challenge

#### Tools for dictionary construction
- Clustering & topic modeling
- Frequency list of words
- Phrase extraction
- Named entity recognition (NER)
- Thesauri & lexical databases
- Identification of inflected forms
- Identification of misspelled words

---

#### THREE MAJOR OBSTACLES

1) Very large number of word forms

2) Polymorphy of language
   - One idea $\rightarrow$ multiple forms

3) Polysemy of words
   - One word $\rightarrow$ many ideas

---

### Text Analytics Challenge

#### Challenge #3 – Polysemy of words

#### Keyword in Context List (KWIC)

**Senses of word “stress”**

#1 (psychology) a state of mental or emotional strain or suspense

#2 (physics) force that produces strain on a physical body

#3 Verb - single out as important
**Challenge #3 – Polysemy of words**

**Keyword in Context List (KWIC)**

Disambiguation using phrases

<table>
<thead>
<tr>
<th>STRESS<em>_THE or STRESS</em>_THAT</th>
<th>“single out as important”</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNDER_STRESS or THEIR_STRESS</td>
<td>Emotional State</td>
</tr>
</tbody>
</table>

**Rule of Thumb**

**PROPOSED BY BENGSTON & XU (1995)**

- Every single item in a dictionary should produce at least 80% of true positives (TP).
- If not, try to remove false positives (FP) using associated phrases until FP is less than 20%.
- If TP still below 80%, remove the word from the dictionary and add associated TP phrases.

**CAUTION:** The 80% criteria do not take into account costs associated with false negatives.

**Challenge #4 – Misspellings**

**1.8 million student comments**

- More than 35 million words
- 78,159 word forms
- 46,404 “unknown” words
  - 75% misspellings (≈ 35,000)
  - 21% proper names (products & people)
  - 4% acronyms
Challenge #4 – Misspellings

61 ways to be “Enthusiastic”

Fuzzy and phonetic string comparison algorithms:
- Damerau-Levenshtein
- Koelner Phonetik
- SoundEx
- Metaphone
- Double-Metaphone
- NGram
- Dice
- Jaro-Winkler
- Needleman-Wunch
- Smith-Waterman-Gotoh
- Monge-Elkan

Automatic Document Classification

1) Training Phase
2) Classification of documents

Automatic Document Classification

- Naïve Bayes Classifiers
  - Probabilistic classifiers that are based on Bayes’ theorem, which states that the probability of an event’s occurrence is equal to the intrinsic probability times the probability that it will happen again (naïve = simplistic assumption that the objects are completely independent of one another)
- Rocchio Classification
- K-Nearest Neighbor Method
  - A method to cluster documents based on their distance to the K nearest “neighbor” documents
- Support Vector Machine

Machine Learning

- Algorithmic approach to text to:
  - Recommendations/Predictions (Pandora/Amazon)
  - Classification (Known data to define new data =spam
  - Clustering (New groups of similar data=Google News)
- Large Data Sets (Large Numbers of Words or Phrases)
  - Bag of Words Approach
  - High-Dimensional Vector Spaces
- Common ML algorithms for text categorization
  - Artificial Neural networks
  - Decision trees
  - Support Vector Machines (SVM)
- Supervised Machine Learning
  - Providing a set of “input features” (e.g. terms) can be provided to help enable Machine Learning (ML)
  - An iterative process, where outputs are compared with known values
- Unsupervised Machine Learning
  - Classification of documents where the categories of a test set are not known