TEXT MINING CHALLENGES AND SOLUTIONS IN BIG DATA

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

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

PROVALIS RESEARCH | PROVEN ANALYTICS PLATFORMS

AMERICAN UNIVERSITY WASHINGTON, DC

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

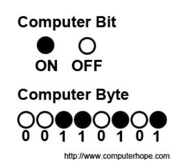
   And if enough time
6. Understand some of the basic features of automatic document classification techniques

Recommended Texts

- *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Gary Miner et al, Academic Press/Elsevier (2012) (Available for Kindle)
- *R for Everyone: Advanced Analytics and Graphics*. Jared P. Lander. Addison-Wesley, 2014. (Available for Kindle)
- *An Introduction to Data Science*, Jeffrey Stanton (2013). (Free iBook or PDF) http://jsresearch.net/wiki/projects/teachdatascience
- *Hadoop: The Engine that Drives Big Data*, Lars Nielsen. Executive Summary (Available for Kindle) New Street Communications (2013)
- *Text Mining for Qualitative Data Analysis in the Social Sciences*. Gregor Wiedemann (2016). Springer.

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



A two bit codebook for response to an invitation

Meaning	2nd Digit	1st Digit
No	0	0
Maybe	0	1
Probably	1	0
Definitely	1	1

Source: Jeff Stanton, Introduction to Data Science

Understanding the "**big**" in Big Data

- Comparison of file sizes:
  - **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 (2^{30}) bytes. 1,024 Megabytes, or 1,048,576 Kilobytes (***894,784 pages of plaintext***)
  - **Terabyte** (TB) 1,099,511,627,776 (2^{40}) bytes, 1,024 Gigabytes, or 1,048,576 Megabytes (***916,259,689 pages of plaintext***)
  - **Petabyte** (PB) 1,125,899,906,842,624 (2^{50}) 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 (2^{60}) bytes, 1,024 Petabytes, 1,048,576 Terabytes, 1,073,741,824 Gigabytes, or 1,099,511,627,776 Megabytes (***960,767,920,505,705 pages of plaintext***)
  - **Zettabyte** (ZB) 1,180,591,620,717,411,303,424 (2^{70}) bytes, 1,024 Exabytes, 1,048,576 Petabytes, 1,073,741,824 Terabytes, 1,099,511,627,776 Gigabytes, or 1,125,899,910,000,000 Megabytes (***983,826,350,597,842,752 pages of plaintext***)
  - **Yottabyte** (YB) 1,208,925,819,614,629,174,706,176 (2^{80}) bytes, 1,024 Zettabytes, 1,048,576 Exabytes, 1,073,741,824 Petabytes, 1,099,511,627,776 Terabytes, 1,125,899,910,000,000 Gigabytes, or 1,152,921,500,000,000,000 Megabytes (***1,007,438,183,012,190,978,921 pages of plaintext***)

Source: http://www.computerhope.com/issues/chspace.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

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

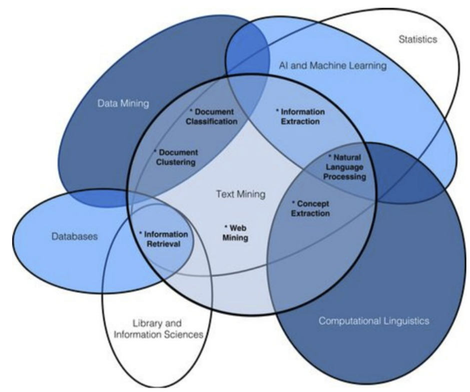
Acquiring Text Data

- Tools and Techniques for scraping sites
  - Software
    - Sitesucker*
    - PageSucker
    - WebGrabber
    - Web Dumper
  - Hand Coding
    - Perl
    - Python
    - Ruby
- Downloading email listserve 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



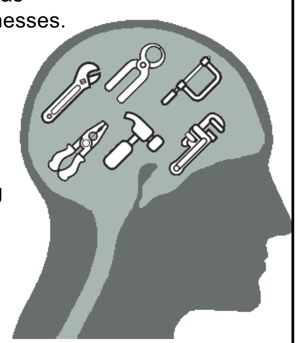
Text Analysis Landscape

FOUR APPROCHES TO TEXT ANALYSIS

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

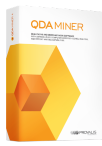
Our Proposal

- 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

Content Analysis & Text Mining



Qualitative Analysis & Mixed Methods

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

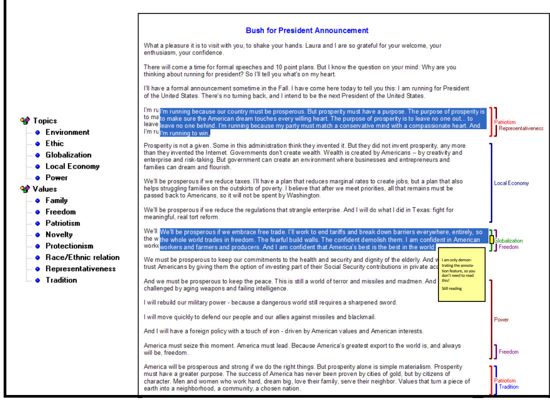
Text mining modules R programming modules (R/tm)
Gensim, Mallet, Quanteda, Rapid Miner, Gate, KNIME

NLP Libraries

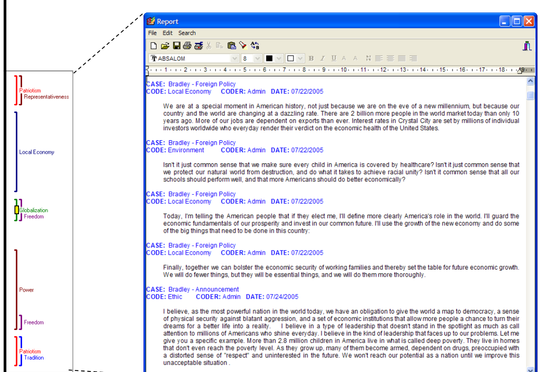
Stanford NLP , Natural Language Tookit (NLTK) OpenCalais
Apache OpenNLP, etc.

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

Computer Assisted Qualitative Analysis



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

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

Challenge #1 – Quantity

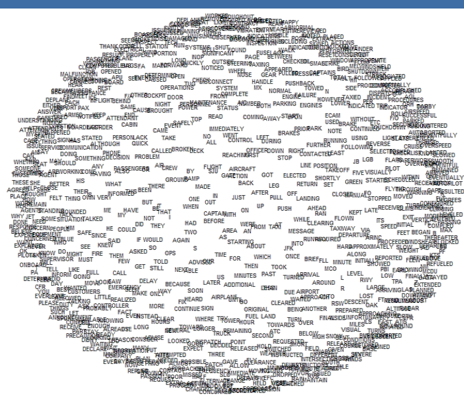
38,996 comments about hotels

- 2,1 million words (tokens)
- 20,116 terms or word forms (types)

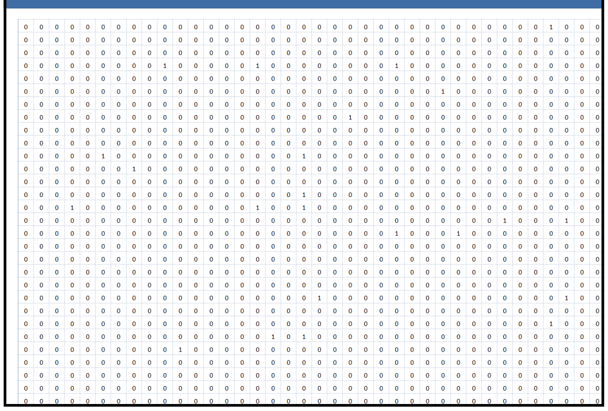
1,8 million course evaluations

- 35 millions words (tokens)
- 78,159 terms or word forms (types)

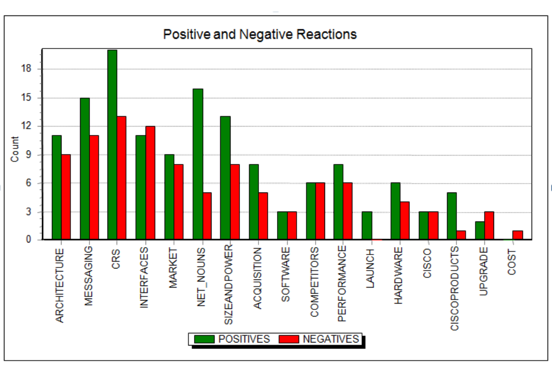
Text Analytics Challenge



Text Analytics Challenge



Text Analytics Challenge



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

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

Challenge #1 – Solutions

Linguistic Pre-processing

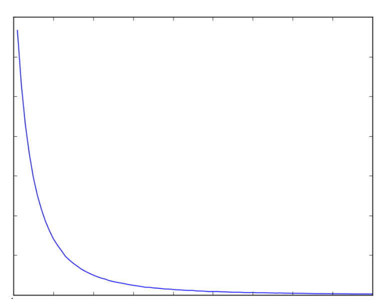
- Removal of stop words
- Stemming
- Lemmatization

Statistical tools

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

The Statistics of Text

Distribution of words: Zipf distribution



The Statistics of Text

38,988 comments about hotels

2.1 M words (20,114 different terms)

MOST FREQUENT TERMS	PERCENTAGE OF TERMS	PERCENTAGE OF WORDS
49	0.24%	50%
300	1.5%	76%
500	2.5%	83%
1000	5.0%	90%

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: $f_{t,d}$ x log (N / n_t)

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

Text Analytics Challenge

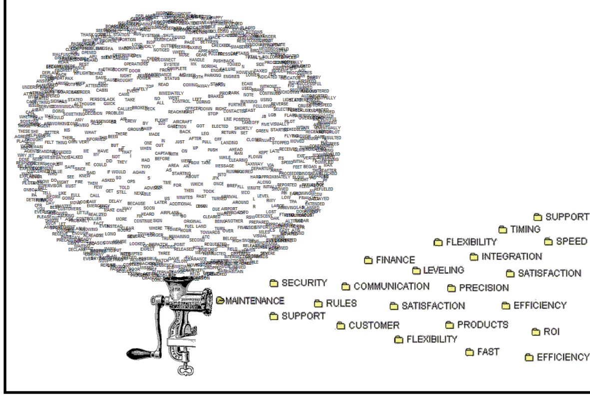
THREE MAJOR OBSTACLES

1) Very large number of word forms

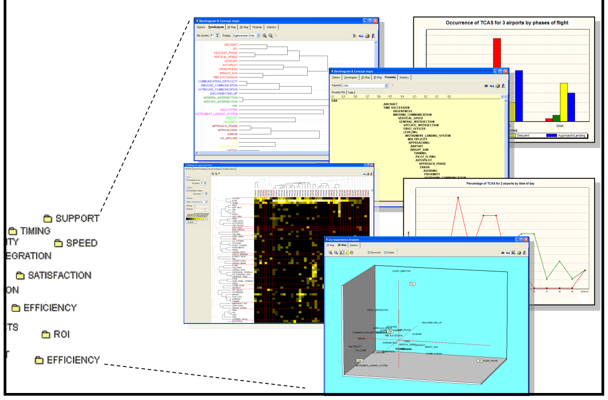
2) Polymorphy of language

One idea → multiple forms

Content analysis method



Content analysis method



Content analysis method

- SENSE_OF_HUMOR
- SOCIALISATION
- CONFLICT_RESOLUTION
- TEAM_WORK
- ORGANIZATION_SKILLS
  - ORGANIZATION_SKILLS (1)
  - ORGANIZATIONAL_SKILLS (1)
  - TIME_MANAGEMENT (1)
  - COMPLETE_A_TASK (1)
- PROBLEM_SOLVING
  - PROBLEM_SOLVE (1)
  - PROBLEM_SOLVE_SKILL (1)
  - PROBLEM_SOLVING (1)
  - PROBLEM_SOLVING_SKILLS (1)
  - SOLVE_A_PROBLEM (1)
  - SOLVE_PROBLEMS (1)
  - SOLVING_PROBLEMS (1)
- LEADERSHIP
  - LEAD_A_GROUP (1)
  - LEAD_A_TEAM (1)
  - LEAD_GROUPS (1)
  - LEAD_PEOPLE (1)
  - LEADERSHIP_SKILLS (1)
  - LEADING_A_GROUP (1)
  - LEADING_A_LARGE (1)
  - LEADING_A_TEAM (1)
  - LEADING_GROUPS (1)
  - LEADING_TEAMS (1)
  - TAKING_CHARGE (1)
  - TAKING_COMMAND (1)
  - TAKING_THE_LEAD (1)
  - GROUP_LEADERSHIP (1)
  - TEAM_BUILDING (1)
  - TEAM_BUILDING_SKILLS (1)
  - GUILD_LEADER (1)
  - GROUP_LEADER (1)
  - GOOD_LEADER (1)
  - TEAM_LEADER (1)
- CREATIVITY
  - CREATIVE_OUTLET (1)
  - CREATIVE_THINKING (1)
  - CREATIVE_SIDE (1)
  - CREATIVE_WRITING (1)
  - CREATIVE_PROCESS (1)
  - CHARACTER_CREATION (1)

- 5- HUMAN FACTORS
  - ASSERTIVENESS
  - AWARENESS
  - ARGUMENTATIVE
  - DISTRACTION
    - DISTRACTED
    - DISTRACTION
    - INATTENTION
    - INATTENTIVE
    - UNALERT
    - UNAWARE
    - UNOBSERVANT
    - UNVIGILANT
    - UNWATCHFUL
  - MISUNDERSTANDING
  - MISTAKE
  - OMISSION
    - DISREGARDED
    - FORGET
    - FORGOTTEN
    - IGNORE
    - IGNORED
    - NEGLECTED
    - NEGLIGENCE
    - OMISSION
    - OMIT
    - OMITTED
    - OVERLOOKED
- POSITIVES
  - BRILLIANT
  - DAZZLING
  - DELIGHTFUL
  - ENABLE
  - EXCELLENT
  - FAST
  - GOOD
  - INTERESTING
  - LIKE_IT
  - LOVE
  - NICE
  - WORKS_FLAWLESS
  - WORTHY
- NEGATIVES
  - ARDUOUS

Custom Dictionary for Course Evaluation

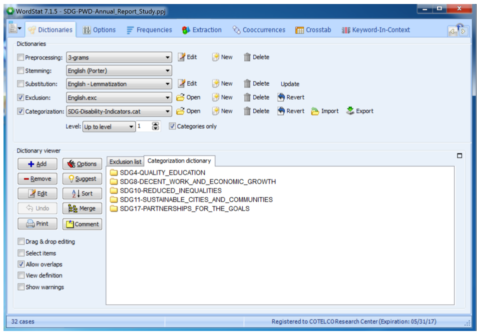
POSITIVE		NEGATIVE	
CLEAR	accurate, explicit, precise, straightforward	UNCLEAR	ambiguous, confusing, vague, lack of clarity
EASY	effortless, manageable, simple, smooth	DIFFICULT	challenging, complex, laborious, tough
COMPREHENSIVE	complete, detailed, exhaustive, great depth	SUPERFICIAL	insipid, lack of depth, shallow, scratched the surface
RELEVANT	essential, pertinent, useful, worthwhile	NOT WORTHWHILE	complete waste, irrelevant, obsolete, pointless
ENGAGING	appealing, captivating, grabs your attention, motivating	NOT ENGAGING	insipid, lacks excitement, lose interest, wearisome
INTERESTING	absorbing, exciting, fascinating, stimulating	BORING	depressing, drowsy, dull, yawn
ENJOYABLE	agreeable, delightful, gratifying, pleasant	FRUSTRATING	annoying, antagonizing, irksome, irritating
FAIR	impartial, integrity, objective, unbiased	UNFAIR	bias, discriminated, favouritism, unethical
HELPFUL / SUPPORTIVE	assists, great assistance, provides assistance, reinforces	UNHELPFUL	not cooperative, not useful, not informative, had zero patience
AVAILABLE	accessible, always available	UNAVAILABLE	inaccessible, lack of assistance, no
RESPONSIVE	answer my questions, quick to reply,	UNRESPONSIVE	no answer, nobody responded, no
APPROACHABLE	easy to talk to, laidback, good natured,	UNAPPROACHABLE	detached, distant, intimidating,

Custom Dictionary for SDGs and PWDs

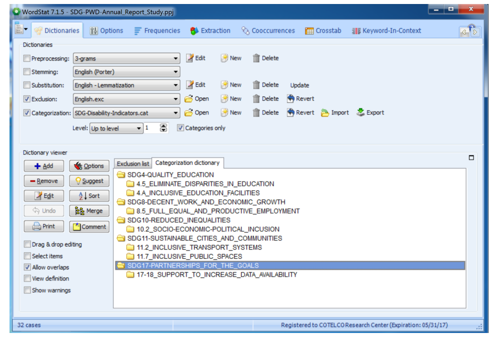
- **SDG Goal 4: Education** = 2
  - *4.5 By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with **disabilities**, indigenous peoples and children in vulnerable situations*
  - *4.a. Build and upgrade education facilities that are child, **disability** and gender **sensitive** and provide safe, nonviolent, inclusive and effective learning environments for all*

- **Goal 4.5:**
  - Gender Disparity
    - Key words
    - Key phrases
  - Equal Access to Education
  - Equal Access to Vocational Training
- **Goal 4.a**
  - Facilities
  - Safe Environment
  - Inclusive Environment
  - Effective Learning

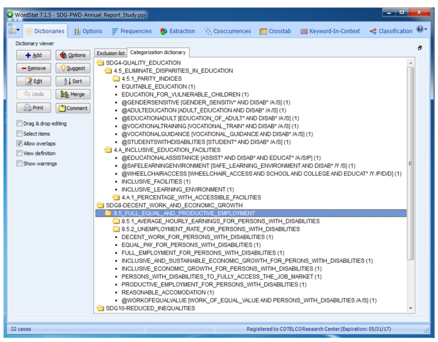
Custom Dictionary for SDGs and PWDs



Custom Dictionary for SDGs and PWDs



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

- DICTION - Hart
- Lasswell Value Dictionary - Lasswell
- General Inquirer Harvard IV - Stone

Measure Latent Dimensions

COMMUNICATION VAGUENESS DICTIONARY

BLUFF AND RECOVERY
- ACTUALLY (1)
- ALRIGHT (1)
- ANYHOW (1)
- ANYWAY (1)
- AS_A_MATTER_OF_FACT (1)
- AS_FAR_AS (1)
- AS_WE_KNOW (1)
- AT_ANY_RATE (1)
- A_LONG_STORY_SHORT (1)

ADMISSION OF ERROR
- A_MISTAKE (1)
- CHANCES_ARE (1)
- I_FORGOT (1)
- I_GUESS (1)
- I_MADE_A_MISTAKE (1)
- I_SLIPPED (1)
- I_SUPPOSE (1)
- I_THINK (1)
- I_WAS_WRONG (1)

INDEFINITE AMOUNT
- A_BIT (1)
- A_BUNCH (1)
- A_COUPLE (1)
- A_LITTLE (1)
- A_LOT (1)
- CONSIDERABLE (1)
- CONSIDERABLE_AMOUNT (1)
- FEW (1)
- LESSER_EXTENT (1)

RESERVATIONS
- APPARENTLY (1)
- BASICALLY (1)
- CONSIDERABLY (1)
- ESSENTIALLY (1)
- FOR_THE_MOST_PART (1)
- IN_ESSENCE (1)
- IN_GENERAL (1)
- IN_SUBSTANCE (1)
- MATTER_OF_DEGREE (1)

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

Text Analytics Challenge

THREE MAJOR OBSTACLES

1) Very large number of word forms

2) Polymorphy of language

One idea → multiple forms

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

Text Analytics Challenge

SEXUAL SCALE		Total 473
False positives		**104**
AIDS	band-aids	2
BI	bi-partisan	10
DICK	Dick Cheney, Dick Lugar	38
HUMP*	Hubert Humphress	1
KISS*	Kissinger	1
VIRGIN*	Virginia, Virginians	51
BOOB*	booby-trapped	1
Questionable positives		**242**
LOVE*	love, loved, loves	208
PASSION*	passion, passions, passionate	26
BREAST*	breast cancer	6
OVAR*	ovarian cancer	2

Challenge #3 – Polysemy of words

Keyword in Context List (KWIC)

RECNO		KEYWORD	
2766	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRAFT AT APPROXIMATELY 1
2690	is still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	is signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	ecovering from brain surgery and wasn't suppose to be under	stress	and wasnt going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	, tired and had very little food. She had taken dramamine and h
7972	d very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we p
8702	F2 - Passenger was under	stress	, tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	Mr.Koch took the flt back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consic

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)

RECNO		KEYWORD	
2766	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRAFT AT APPROXIMATELY 1
2690	ıs still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	ıs signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	ecovering from brain surgery and wasn't suppose to be under	stress	and wasnt going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	, tired and had very little food. She had taken dramamine and I
7972	d very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we p
8702	F2 - Passenger was under	stress	, tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	Mr.Koch took the flt back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consic

Disambiguation using phrases

STRESS*_THE **or** STRESS*_THAT → "single out as important"

UNDER_STRESS **or** THEIR STRESS → Emotional State

Item Matching Rules

IMPROPER MATCHING RULES

- Any matching items
- First item encountered in alphabetically sorted list

PROPER MATCHING PRIORITY RULES

- First item encountered in a carefully arranged list or
- Longer phrases over shorter phrases
- Phrases over words
- Words over word patterns
- Longer word patterns over shorter ones

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 it 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.

Rule of Thumb

United States Department of Agriculture

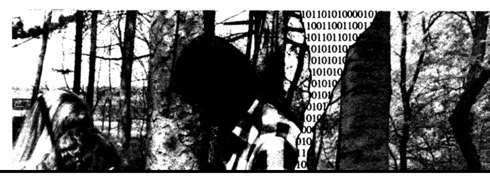
Forest Service

North Central Forest Experiment Station

Research Paper NC-323

Changing National Forest Values: A Content Analysis

David N. Bengston and Zhi Xu



Challenge #3 – Polysemy of words

Keyword in Context List (KWIC)

RECNO		KEYWORD	
2766	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRAFT AT APPROXIMATELY 1
2690	ıs still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	ıs signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	ecovering from brain surgery and wasn't suppose to be under	stress	and wasnt going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	, tired and had very little food. She had taken dramamine and I
7972	d very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we p
8702	F2 - Passenger was under	stress	, tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	Mr.Koch took the flt back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consic

Disambiguation using rules

TRANSFER* IS NEAR **TECHNOLOGY**

TRANSFER* IS NOT NEAR **BUS**

SATISFIED IS AFTER **#NEGATION**

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"

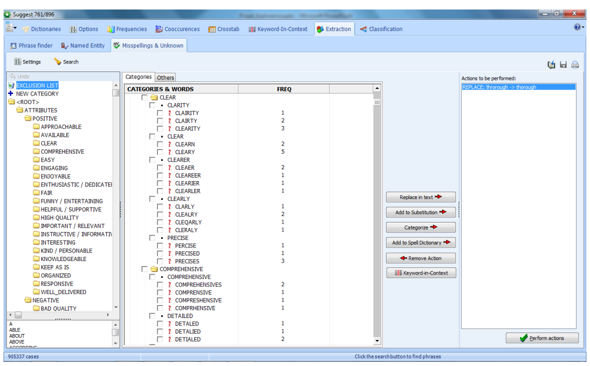
EHNTHUSIASTIC	2	ENTHUSAIASTIC	1	ENTHUSICASITC	1
ENHTUSIASTIC	1	ENTHUSAISTIC	8	ENTHUSICASTIC	1
ENTHEUSIASTIC	1	ENTHUSAITIC	1	ENTHUSICATIC	3
ENTHHUSIATIC	1	ENTHUSASITIC	1	ENTHUSISASTIC	1
ENTHIASTIC	1	ENTHUSASTIC	52	ENTHUSISATIC	2
ENTHISIASTIC	2	ENTHUSATIC	11	ENTHUSISTIC	7
ENTHOUSIASTIC	13	ENTHUSIACTIC	4	ENTHUSTATIC	2
ENTHSIASTIC	2	ENTHUSIADTOC	3	ENTHUSTIASTIC	17
ENTHSUASTIC	1	ENTHUSIAITIC	1	ENTHUSUASTIC	4
ENTHUAISTIC	1	ENTHUSIANSTIC	3	ENTHUTIASTIC	2
ENTHUASASTIC	1	ENTHUSIASITC	9	ENTTHUSIASTIC	1
ENTHUASIASTIC	2	ENTHUSIASITIC	5	ENTUISASTIC	1
ENTHUASIATIC	2	ENTHUSIASTC	5	ENTUSIASITC	1
ENTHUASISTIC	1	ENTHUSIASTCI	1	ENTUSIASTHIC	2
ENTHUASTIC	30	ENTHUSIASTICE	2	ENTUSIASTIC	47
ENTHUDIASTIC	1	ENTHUSIASTICS	2	ENTUSIATIC	1
ENTHUIASTIC	20	ENTHUSIASTTIC	1	ENUTHUSIASTIC	1
ENTHUISASTIC	2	ENTHUSIATHIC	1	EUNTHUSIASTIC	1
ENTHUISIASTIC	3	ENTHUSIATIC	185	ANTHUSIASTIC	1
ENTHUSIATSTIC	3	ENTHUSIATSIC	4	ENTUSIASIC	1
				ETHUSIASTIC	28

Challenge #4 – Misspellings

Fuzzy and phonetic string comparison algorithms:

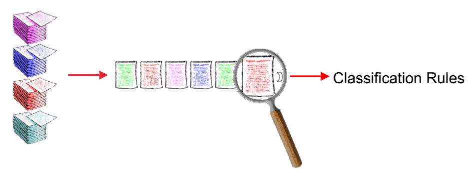
- Damerau-Levenshtein
- Koelner Phonetik
- SoundEx
- Metaphone
- Double-Metaphone
- NGram
- Dice
- Jaro-Winkler
- Needleman-Wunch
- Smith-Waterman-Gotoh
- Monge-Elkan

Challenge #4 – Misspellings

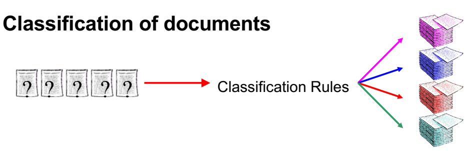


Automatic Document Classification

1) Training Phase

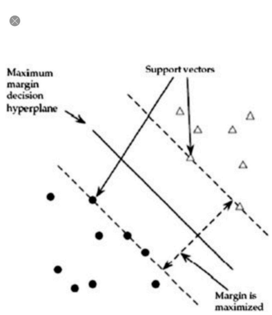


2) Classification of documents



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