Text Mining Tools for Qualitative Researchers: A Curse or a boon?

Normand Péladeau
President
Provalis Research Corp.
peladeau@provalisresearch.com
ANALYSIS OF TEXTUAL DATA

- Qualitative Researchers
- Market Researchers
- Pollsters
- Journalists
- Historians
- Archivists
- Librarians
- Lawyers and Paralegal Professionals
- Crime Analysts
What are all those people trying to achieve?

- Accurately describe a situation
- Find communalities and differences
- Find hidden patterns and relationships
- Retrieve relevant information
- Generate new knowledge or discovery
- Generate and test hypothesis
- Etc.
Which technique do they use?

ANALYSIS OF TEXTUAL DATA

- Qualitative Analysis
- Content Analysis
- Text Mining
- Information Retrieval
- Computational Linguistics
- Knowledge Management
The Landscape of Text Analysis Tools

- **Qualitative Analysis**: Manual reading and coding of documents
- **Content Analysis**
- **Text Mining**
The Landscape of Text Analysis Tools

- **Qualitative Analysis**: Manual reading and coding of documents
- **Content Analysis**: Dictionaries of words, phrases, patterns, rules
- **Text Mining**
The Landscape of Text Analysis Tools

- **Qualitative Analysis**: Manual reading and coding of documents
- **Content Analysis**: Dictionaries of words, phrases, patterns, rules
- **Text Mining**: Statistical analysis, NLP and data mining techniques
Text Mining approach
The Landscape of Text Analysis Tools

- **Qualitative Analysis**: Manual reading and coding of documents
- **Content Analysis**: Dictionaries of words, phrases, patterns, rules
- **Text Mining**: Statistical analysis, NLP and data mining techniques

QUAL

CATA
The Landscape of Text Analysis Tools

Qualitative Analysis
- Atlas.ti
- Nvivo
- MaxQDA
- Qualrus
- Ethnograph
- HyperResearch
- Dedoose
- QDA Miner
- Alceste
- Clarabridge
- SAS Text Miner
- Catpac
- Leximancer
- T-Lab
- Lexiquest
- WordStat

Content Analysis
- General Inquirer
- Diction
- LIWC
- Tabari
- TextQuest
- TextPack
- Yoshikoder

Text Mining
- Alceste
- Clarabridge
- SAS Text Miner
- Catpac
- Leximancer
- T-Lab
- Lexiquest
- WordStat

PROVALIS RESEARCH | PROVEN ANALYTICS PLATFORMS
Mutual Contempt

FOR QUALITATIVE RESEARCHERS

• Counting words is meaningless
• Computers cannot replace human judgment
• Scepticism toward forms of computer assistance or automation

FOR QUANTITATIVE TEXT ANALYSTS

• Human coding it too time consuming and does not scale up
• Human coding is too unreliable and subjective
• For some, computer coding can replace human coders
Various typologies in mixed methods

QUAL + quan
QUAL → quan
QUAN + qual
QUAN → qual
QUAL (quan)
QUAN (qual)
etc.

Triangulation of QUAL and QUAN results
Exploratory use of both QUAL and QUAN
Explanatory use of QUAL for QUAN
Confirmatory use of QUAN for QUAL
etc.
Various typologies in mixed methods

QUAL + **cata**
QUAL \( \rightarrow \) **cata**
**CATA** + qual
**CATA** \( \rightarrow \) qual
QUAL (**cata**)  
**CATA** (qual)
etc.

Triangulation of QUAL and **CATA** results
Exploratory use of both QUAL and **CATA**
Explanatory use of QUAL for **CATA**
Confirmatory use of **CATA** for QUAL
etc.
Potential Benefits of CATA to QDA

- Improve the sampling process
- Perform data reduction
- Speed up familiarisation with the text data
- Assist the structuring of the codebook
- Speed up / automate the coding process
- Increase the reliability of the coding process
- Increase the generalizability of the conclusions
**TASK:** Analyse a limited number of documents from a large collection.

**SAMPLING OBJECTIVE:**
Select documents that are

- representative of the points of view of the majority
- sensitive to alternate points of view
CLIENT: Berezowski, Snyder, & Mclarty (2008)
Alberta Agriculture - Food And Rural Development

TASK: Classification of veterinarian records for real time surveillance

DATA: 35,720 cattle testing reports
- clinical signs and presumptive diagnosis in free text format
- technical & non-technical terms, misspellings, etc.

OBJECTIVES:
• Identify potential “Clinical Suspects” of major health risks
• Classify submissions into clinical syndromes
Data Reduction

Sample Dictionary Entries

BEHAVIORAL
- APPREHENS*
- AVOID*
- BALKING
- BAWLING
- BELIGER*
- BELLIGER
- BIZARRE
- COMPULSIVELY
- CRAZY
- DELIRIOUS
- DELUSIONAL
- Demeanor
- DEMENTIA
- DERANGED
- DOPEY
- DRUNK
- ENRAGED
- ERRATIC*
- EVASIVE
- FLIGHTY
- FRACTIOUS
- FRENZY
- GAZING
- GROUCHY
- HEAD-SHY
- HIDING
- HYSTERIA
- INCOHERENT
- INTOLERAN*
- MEAN
- MENTAL*

CHRONIC
- ADVANCED
- AWHILE
- CACHEXI*
- CACHEZIA
- CHACHAXIA
- CHACHTIC
- CHACHEXIA
- CHONIC
- CHINIC
- CHRONIC
- CHRONI*
- CRONIC
- CURABLE
- D*BLIT*
- DAYS_AGO
- DEBILITATED
- DEBILITATED
- DECLIN*
- DEGENERAT*
- DEGENERATIVE
- DEGNERAT*
- DEBILITATING
- DOWNHIL
- DOWNHIL*
- DURATION
- EMACATION
- EMACIAT*
- EMACIAT*
- EMACICA*
- FAIL*
- GRADLUAI*

NEUROLOGICAL
- *TILT*
- CONVULS*
- EPISTHOTONIS
- FLACCID
- HYPERATHESIA
- KNUCKLE
- N*URO*
- NEUR*
- OPINOTHATONES
- OPISTHITANUS
- OPISTHOT*
- OPISTLOTONES
- OPITHOLONUS
- OPITHOTONUS
- OPOTHOTENUS
- OPOTHATONUS
- PADDEL
- PADDDLING
- PARALY*
- PARAPARESIS
- PAREISIS
- PARES*
- PARETIC
- PARALYSIS
- PARLYSIS
- PARLYZED
- PARYALYZED
- PROPRI*
- SEIZUR*
- TREMOR*
- WRITHING

LOCROMOTOR
- *NCOOD*
- *NCOORD*
- *NCO-ORDINAT*
- *OCOM*
- BUCKLING
- CLUMSY
- CO-ORDINAT*
- COORDINATE*
- CYCLIC
- CYCLING
- DIZZY

ATAXIA
- *ATAK*
- AT*XIA
- AT*XIC
- ATACHIA
- ATAXIA
- TAXIA

HYPEREXCITABLE
- ANXIETY
- ANXIOUS
- CHARG*
- CHASE*
- EXCITABLE
- HYPERALERT
- HYPEREXC*
- HYPEREXCITABLE
- HYPERSENSITIV*
- IRRITA*
## Data Reduction

**Data reduction process**

**Clinical Suspects**

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Submissions</td>
<td>35,721</td>
</tr>
<tr>
<td>Neuro + Behavior</td>
<td>4,583</td>
</tr>
<tr>
<td>Rule Outs</td>
<td>4,010</td>
</tr>
<tr>
<td>Clinical Suspects</td>
<td>573</td>
</tr>
</tbody>
</table>
**TASK:** Create a codebook of topics mentioned in a large text collection

“In principles we could organize the data by grouping like with like [...] We can put all the bits of data which seem similar or related into separate piles, and then compare the bits within each pile. We may even want to divide up the items into a pile into separate ‘sub-piles’ if the data merits further differentiation”

(Dey, 1993, p.95)
parent education, after school programmes
parenting education made compulsory at school
Education in communities, schools etc
keeping children entertained and active after school and on weekends
Safe Havens, school counsellors, school initiatives, contraception
Extensive education in schools on bringing up children.

parenting skills and support
Courses on parenting skills for parents
Parenting skills programmes for all.
Helping Young parents in parenting skills
Parenting skills for young, as well as new, parents.

drug and alcohol abuse
Alcohol and drug prohibition
drug and alcohol abuse
Drug and alcohol abuse. Reintroduce six o'clock closing.
alcohol drug and other agencies to work with families and the addicted
Education, Parenting programmes, Social services, Drug & Alcohol etc
Education, with an emphasis on drug and alcohol use and abuse
more staff in hospitals, police, social workers
Police and social workers
More community midwifery and social worker input.
more frontline staff eg social workers, police youth aid etc.

Incomes for low income families help those on low incomes more
More money to low income families...
Community based Agencies that support low-income families
Fund community agencies who offer support to low-income families
Low income families need more income and this creates pressure.
some agency supporting low income
Families with low incomes

Fund families to look after each other
Fund healthy parenting courses
Funding in schools Funding in hospitals Funding in poor neighborhood
education and funding for help centers
Funding of organizations like Parent Inc to help them help more people
### Cluster Codes

<table>
<thead>
<tr>
<th>Cluster Codes</th>
<th>Similarity</th>
<th>Clustered Items</th>
<th>Item Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOOD INTERNET</td>
<td>2.55</td>
<td>Only complaint is that a resort such as this can and should have WIRELESS INTERNET access by now.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.58</td>
<td>Except for one small area on the first floor, there is no WIRELESS INTERNET.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.60</td>
<td>One downside was the lack of WIRELESS INTERNET.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.61</td>
<td>They have INTERNET WIRELESS but it is hit and miss.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.63</td>
<td>Only to find that there is NO WIRELESS INTERNET in the building.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.64</td>
<td>Everything was really good except there was no WIRELESS INTERNET in the rooms.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.65</td>
<td>Good stay and value but room INTERNET &amp; WIRED not WIRELESS and did not work.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.67</td>
<td>INTERNET never worked no WIRELESS.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.71</td>
<td>Rooms don't have WIRELESS INTERNET.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.76</td>
<td>No WIRELESS INTERNET in hotel at all.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.79</td>
<td>No WIRELESS INTERNET other.</td>
<td></td>
</tr>
<tr>
<td>FEES NEGATIVE</td>
<td>2.58</td>
<td>Be prepared for those HIDDEN FEES, the rooms are old and they have a bad smell.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.59</td>
<td>HIDDEN FEES, rude bell staff, nice pool.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.70</td>
<td>No HIDDEN FEES.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.76</td>
<td>HIDDEN FEES.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.78</td>
<td>I'm not sure what the &quot;Resort Fee&quot; got us and there seemed to be a lot of other HIDDEN FEES.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.80</td>
<td>The other good thing is they don't rip you off with HIDDEN resort FEES.</td>
<td></td>
</tr>
<tr>
<td>RESORT FEE</td>
<td>3.69</td>
<td>The stay was great but I was disappointed to get an ADDITIONAL charge of about $90 for RESORT FEE.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.63</td>
<td>Very disappointed that an ADDITIONAL charge (RESORT FEE) was added to the daily room price.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.72</td>
<td>Also, when they say there is an ADDITIONAL 'RESORT FEE' for $13, it is actually for $20.12.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.88</td>
<td>NO SAFES in ROOM.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.70</td>
<td>The only problem NO SAFES in the hotel ROOM.</td>
<td></td>
</tr>
</tbody>
</table>

### Item Codes
- WIRELESS INTERNET
- RESORT FEE
- FEES
- HIDDEN FEES
- SAFES in ROOM

---

**Clustered Items:**

- FOOD INTERNET
- INTERNET
- FEES NEGATIVE
- RESORT FEE

**Unclassified Items:**

- 2.22 ONLY COMPLAINT WOULD BE THE RESORT FEE WHICH MYSELF AND YELLOW TRAVELLER FELT WAS UNACCEPTABLE, FEEL THAT IS NOT VALUE
- 2.25 Great stay, all the amenities at our hand, pedestrian to different resorts gives you the opportunity to visit without even going outside, on look walk, no WIRE INTERNET, only WIRED INTERNET available in the room.
- 2.25 We stayed in an "upgraded" room and it did not have wifi (still hard WIRED INTERNET throughout the hotel), light bulbs were out, hair dryer was broken, etc.
- 2.25 There's only WIRED INTERNET available, and if you have a newer laptop where you can't connect ethernet cables then forget it.
- 2.24 The only problem was the lack of wifi, these days with Pad and smart technology, WIRED INTERNET is old fashioned and useless.
- 2.23 It was expensive and said that it covered free omelette, free newspaper and WIRED INTERNET.
- 2.22 It was expensive and said that it covered free omelette, free newspaper and WIRED INTERNET.
- 2.22 The location is decent and the resort fees are considerably less than other properties with WIRED INTERNET and room service included.
Clustering of Words

Small Clusters

CHILD
YOUTH
ALCOHOL
DRUG
INCOME
LOW
EDUCATION
PARENT
PARENTAL
RESPONSIBILITY
FRONTLINE
STAFF
FRONT
LINE
SOCIAL
WORKER
HOME
VISIT
Clustering of Words

Larger Clusters

- Abuse
- Stop
- Child
- Youth
- Alcohol
- Drug
- Programmes
  - Agency
  - Identify
  - Risk
  - Family
- Support
- Income
- Low
- Education
- Parent
- People
- Young
- Base
- Community
- Program
- Early
- Intervention
- Service
- Prevention
- Contraception
- Free
- Make
- Children
- Remove
- Offender
Clustering of Words

Even Larger Clusters
established by our Founding Fathers. The Supreme Court and all federal courts need to get back to the business of interpreting the Constitution and the law, and out of appointed leaders, not anointed judges, whether in the nation, or right here in the state of New Hampshire.

A rollback of federal power, fewer federal regulations and lower taxes, welfare, poverty, housing, primary and secondary education should be returned to the states. A simple, fair, tax system every American can understand.

books in Washington. America's greatest deficit is no longer for national defense, but is found within the soul, and it may be found in a polluted and poisoned culture that pervades America from coast to coast. A culture of death, a culture of death, a culture of death. As long as I have breathe in me, there will be at least one and only one party in the United States that dares, without apology, to stand up for the rights of the unborn, the elderly, the sick, the handicapped. This is the Great Generation that brought us through World War II and through the Cold War. This is the Great Generation that brought through the Depression. I can yet recall, as a boy, seeing my father go off to work three jobs, to provide food for my family. I can recall the four younger brothers stopping by the house on Chestnut Street to say how their brothers contributed. It was our generation, all of us, who stayed the course, persevered, and prevailed in the long twilight struggle against one of the most monstrous
<table>
<thead>
<tr>
<th>Text</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirty, dirty, dirty</td>
<td>4.201</td>
</tr>
<tr>
<td>1st room was very dirty, and the toilet didn't work!!! 2nd room was better some of the service desk people we're super rude!</td>
<td>3.641</td>
</tr>
<tr>
<td>Dirty room and service is just so so.</td>
<td>2.697</td>
</tr>
<tr>
<td>Our room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Our room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Your room was very dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>room was dirty            .</td>
<td>2.510</td>
</tr>
<tr>
<td>Dirty</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>The hotel overall is also very dirty.</td>
<td>1.939</td>
</tr>
<tr>
<td>The hotel was dirty.</td>
<td>1.939</td>
</tr>
<tr>
<td>The hotel was dirty, the rooms where dirty, my friend didn't get his room cleaned for three days.</td>
<td>1.533</td>
</tr>
<tr>
<td>It was a little better but still dirty.</td>
<td>1.493</td>
</tr>
<tr>
<td>Very dirty room and bathroom.</td>
<td>1.431</td>
</tr>
<tr>
<td>The room is dirty; the desk clerk is rude.</td>
<td>1.329</td>
</tr>
<tr>
<td>This place was very dirty.</td>
<td>1.297</td>
</tr>
<tr>
<td>All service was bad Dirty rooms</td>
<td>1.256</td>
</tr>
<tr>
<td>The room is nice, but a little dirty.</td>
<td>1.168</td>
</tr>
<tr>
<td>It was old and dirty.</td>
<td>1.158</td>
</tr>
<tr>
<td>The carpet in the room was very dirty.</td>
<td>1.153</td>
</tr>
<tr>
<td>Bathroom was dirty.</td>
<td>1.131</td>
</tr>
<tr>
<td>Text</td>
<td>Similarity</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Dirty, dirty, dirty</td>
<td>4.201</td>
</tr>
<tr>
<td>1st room was very dirty, and the toilet didn’t work!! 2nd room was</td>
<td>3.641</td>
</tr>
<tr>
<td>better some of the service desk people were super rude!</td>
<td></td>
</tr>
<tr>
<td>Dirty room and service is just so so.</td>
<td>2.697</td>
</tr>
<tr>
<td>Our room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Our room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Your room was very dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>room was dirty.</td>
<td>2.510</td>
</tr>
<tr>
<td>Dirty</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>Dirty.</td>
<td>2.069</td>
</tr>
<tr>
<td>The hotel overall is also very dirty.</td>
<td>1.939</td>
</tr>
<tr>
<td>The hotel was dirty.</td>
<td>1.939</td>
</tr>
<tr>
<td>The hotel was dirty, the rooms where dirty, my friend didn’t get</td>
<td>1.533</td>
</tr>
<tr>
<td>his room cleaned for three days.</td>
<td></td>
</tr>
<tr>
<td>It was a little better but still dirty.</td>
<td>1.493</td>
</tr>
<tr>
<td>Very dirty room and bathroom.</td>
<td>1.431</td>
</tr>
<tr>
<td>The room is dirty; the desk clerk is rude.</td>
<td>1.329</td>
</tr>
<tr>
<td>This place was very dirty.</td>
<td>1.297</td>
</tr>
<tr>
<td>All service was bad Dirty rooms</td>
<td>1.256</td>
</tr>
<tr>
<td>The room is nice, but a little dirty.</td>
<td>1.168</td>
</tr>
<tr>
<td>It was old and dirty.</td>
<td>1.158</td>
</tr>
<tr>
<td>The carpet in the room was very dirty.</td>
<td>1.153</td>
</tr>
<tr>
<td>Bathroom was dirty.</td>
<td>1.131</td>
</tr>
</tbody>
</table>
My mother was as exuberant as my father was reserved. She and my father married late in life, and I was their only child. I was also her greatest project.
Faster Coding

Query by Example

Retrieval
- Unit: Paragraphs
- In: [SPEEC]

Starting Example:
- Sample text
- Coded segments
- Other project

Examples: [Tradition]
Non examples: [Family;Patriotism]

Options
- Fuzzy string matching
- Do not retrieve already tagged segments
- Minimum criteria: 0.001
- Maximum suggestions: 200
- Language: English
- Add variables: [ ]

Search
Usefullness of Query by Example

FOR BARELY CODED PROJECTS

• Allows to quickly preview the expression of similar ideas
• Allows to immediately code similar ideas across all texts

FOR PARTLY CODED PROJECTS

• Allows to use existing codings to retrieve potentially similar text segments in uncoded documents

FOR FULLY CODED PROJECTS

• Allows to identify potentially false positive (coded segments that should have been coded)

FOR BARELY CODED PROJECTS

• Allows to quickly preview the expression of similar ideas
• Allows to immediately code similar ideas across all texts

FOR PARTLY CODED PROJECTS

• Allows to use existing codings to retrieve potentially similar text segments in uncoded documents

FOR FULLY CODED PROJECTS

• Allows to identify potentially false positive (coded segments that should have been coded)
AUTHOR: Mike Evans (Department of Government and Politics, University of Maryland)

TEXT COLLECTION: Work of Alexander Hamilton (more than 1200 documents & 3 million words)

TASK: Identification of segments where the “master-slave” language used in a metaphorical sense (not literal sense).
STRATEGY:

- Step #1 - Search for SLAVE* and ENSLAVE* (got 47 paragraph).
- Step #2 - Code segments as “Literal” (17) or “Metaphorical” (35).
- Step #3 - Call QUERY BY EXAMPLE.

  EXAMPLES: segments coded as “Metaphorical”
  NON-EXAMPLES: segments coded as “Literal”

  and click SEARCH button.

- Step #4 - Select a few relevant hints, then click SEARCH AGAIN
- Step #5 - Repeat step #4 a couple of times

RESULTS:

- Ended up with 79 relevant segments
- None of the new segments had words matching SLAVE* or ENSLAVE*
Bush for President Announcement

What a pleasure it is to visit with you, to shake your hands. Laura and I are so grateful for your welcome, your enthusiasm, your confidence.

There will come a time for formal speeches and 10 point plans. But I know the question on your mind: Why are you thinking about running for president? So I'll tell you what's on my heart.

I'll have a formal announcement sometime in the Fall. I have come here today to tell you this. I am running for President of the United States. There's no turning back, and I intend to be the next President of the United States.

I'm running because our country must be prosperous. But prosperity must have a purpose. The purpose of prosperity is to make sure the American dream touches every willing heart. The purpose of prosperity is to leave no one out... to leave no one behind. I'm running because my party must match a conservative mind with a compassionate heart. And I'm running to win.

Prosperity is not a given. Some in this administration think they invented it. But they did not invent prosperity, any more than they invented the Internet. Governments don't create wealth. Wealth is created by Americans -- by creativity and enterprise and risk-taking. But government can create an environment where businesses and entrepreneurs and families can dream and flourish.

We'll be prosperous if we reduce taxes. I'll have a plan that reduces marginal rates to create jobs, but a plan that also helps struggling families on the outskirts of poverty. I believe that after we meet priorities, all that remains must be passed back to Americans, so it will not be spent by Washington.

We'll be prosperous if we reduce the regulations that strangle enterprise. And I will do what I did in Texas: fight for meaningful, real tort reform.

We'll be prosperous if we embrace free trade. I'll work to end tariffs and break down barriers everywhere, entirely, so the whole world trades in freedom. The fearful build walls. The confident demolish them. I am confident in American workers and farmers and producers. And I am confident that America's best is the best in the world.

We must be prosperous to keep our commitments to the health and security and dignity of the elderly. And we should trust Americans by giving them the option of investing part of their Social Security contributions in private accounts.

And we must be prosperous to keep the peace. This is still a world of terror and missiles and madmen. And we are challenged by aging weapons and failing intelligence.

I will rebuild our military power - because a dangerous world still requires a sharpened sword.

I will move quickly to defend our people and our allies against missiles and blackmail.

And I will have a foreign policy with a touch of iron - driven by American values and American interests.

America must seize this moment. America must lead. Because America's greatest export to the world is, and always will be, freedom.

America will be prosperous and strong if we do the right things. But prosperity alone is simple materialism. Prosperity must have a greater purpose. The success of America has never been proven by cities of gold, but by citizens of character. Men and women who work hard, dream big, love their family, serve their neighbor. Values that turn a piece of earth into a neighborhood, a community, a chosen nation.
Automation of Coding

- ECONOMIC
  - GLOBALIZATION
    - EXCHANGE (1)
    - EXPORT (1)
    - FREE TRADE (1)
    - GATT (1)
    - GLOBALIZATION (1)
    - IMPORT (1)
    - NAFTA (1)
    - NAFTA (1)
    - TRADE (1)
    - TRADING (1)
    - TRANSPORT (1)
    - WORD TRADE (1)
    - WTO (1)
  - LOCAL ECONOMY
    - BUSINESS (1)
    - CAREER (1)
    - EMPLOY (1)
    - EMPLOYEE (1)
    - ENTERPRISE (1)
    - HOUSING (1)
    - INDUSTRY (1)
    - JOB CREATION (1)
    - JOBS (1)
    - LABOR (1)
    - UNEMPLOYMENT (1)
    - WORK (1)
- OTHER
  - ENVIRONMENT
    - CLEAN ENERGY (1)
    - ENVIRONMENT (1)
    - GREEN (1)
    - PARKS (1)
    - POLLUTION (1)
    - RIVER (1)
  - FREEDOM
    - CHOICE (1)
    - DEMOCRACY (1)
    - FREE (1)
    - FREEDOM (1)
    - INDEPENDENCE (1)
    - LIBERATE (1)
    - LIBERTY (1)
    - TOLERANCE (1)
  - RACE ETHNIC RELATION
    - ABORIGINAL (1)
    - CITIZEN (1)
Automatic Document Classification

1) Training Phase

2) Classification of documents
Automation of Coding

Bush for President Announcement

What a pleasure it is to visit with you, to shake your hands. Laura and I are so grateful for your welcome, your enthusiasm, your confidence.

There will come a time for formal speeches and 10 point plans. But I know the question on your mind: Why are you thinking about running for president? So I'll tell you what's on my heart.

I'll have a formal announcement sometime in the Fall. I have come here today to tell you this. I am running for President of the United States. There’s no turning back, and I intend to be the next President of the United States.

I’m running because our country must be prosperous. But prosperity must have a purpose. The purpose of prosperity is to make sure the American dream touches every willing heart. The purpose of prosperity is to leave no one out... to leave no one behind. I’m running because my party must match a conservative mind with a compassionate heart. And I’m running to win.

Prosperity is not a given. Some in this administration think they invented it. But they did not invent prosperity, any more than they invented the Internet. Governments don’t create wealth. Wealth is created by Americans — by creativity and enterprise and risk-taking. But government can create an environment where businesses and entrepreneurs and families can dream and flourish.

We’ll be prosperous if we reduce taxes. I’ll have a plan that reduces marginal rates to create jobs, but a plan that also helps struggling families on the outskirts of poverty. I believe that after we meet priorities, all that remains must be passed back to Americans, so it will not be spent by Washington.

We’ll be prosperous if we reduce the regulations that strangle enterprise. And I will do what I did in Texas: fight for meaningful, real tort reform.

We’ll be prosperous if we embrace free trade. I’ll work to end tariffs and break down barriers everywhere, entirely, so the whole world trades in freedom. The fearful build walls. The confident demolish them. I am confident in American workers and farmers and producers. And I am confident that America’s best is the best in the world.

We must be prosperous to keep our commitments to the health and security and dignity of the elderly. And we should trust Americans by giving them the option of investing part of their Social Security contributions in private accounts.

And we must be prosperous to keep the peace. This is still a world of terror and missiles and madmen. And we are challenged by aging weapons and failing intelligence.

I will rebuild our military power - because a dangerous world still requires a sharpened sword.

I will move quickly to defend our people and our allies against missiles and blackmail.

And I will have a foreign policy with a touch of iron - driven by American values and American interests.

America must seize this moment. America must lead. Because America’s greatest export to the world is, and always will be, freedom.

America will be prosperous and strong if we do the right things. But prosperity alone is simple materialism. Prosperity must have a greater purpose. The success of America has never been proven by cities of gold, but by citizens of character. Men and women who work hard, dream big, love their family, serve their neighbor. Values that turn a piece of earth into a neighborhood, a community, a chosen nation.
Automation of Coding

- Petroleum
- Representativeness
- Local Economy
- Globalization
- Freedom
- Power
- Tradition

Classification Rules

- What a pleasure it is to have enthusiasm, you know.
- There will come a day when we all think about running for office.
- I'll have a formal dinner for the President of the United States.
- I'm running because I need to make sure that we leave no one behind.
- I'm running to write history.
- Prosperity is not just for the rich; it's also for the poor and working families.
- We'll be prosperous and it will help us remember the Golden Age.
- We'll be prosperous because we've been fighting back for America.
- We must be prosperous because we understand and respect our tradition.
- And we must be prosperous because we must be challenged by agriculture.
- I will rebuild our wealth.
- I will move quickly.
- And I will have a great administration.
- America must see that we will be, freedom.
- America will be great, but our greatness must have a great character. Men and women of earth into a nation.
PSYCHOMETRIC MEASUREMENT

- Linguistic Inquiry and Word Count (LIWC) - Pennebaker
- Regressive Imagery Dictionary (RID) – Martindale
- Communication Vagueness Dictionary – Hiller
- Others

SOCIO-POLITICAL MEASUREMENT

- DICTION
- Lasswell Value Dictionary
- General Inquirer
### TABLE 11.2

Overview of Nine Influential Quantitative Text Analysis Approaches in Psychology

<table>
<thead>
<tr>
<th>Name</th>
<th>Reference</th>
<th>Linguistic parameters</th>
<th>Coding</th>
<th>Conceptual Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thematic Content Analysis</td>
<td>Smith (1992)</td>
<td>Need for power, need for achievement, need for affiliation, explanatory style, integrative complexity</td>
<td>judges</td>
<td>instrumental thematic specific content/style</td>
</tr>
<tr>
<td>Gottschalk–Gieser Method</td>
<td>Gottschalk et al. (1969)</td>
<td>Clinical phenomena (e.g., anxiety, hostility, social alienation, depression, cognitive impairment)</td>
<td>judges/computer</td>
<td>instrumental semantic specific content</td>
</tr>
<tr>
<td>Regressive Imagery Dictionary</td>
<td>Martindale (1990)</td>
<td>29 categories of primary process cognition (e.g., oral, sex, Icarian imagery), 7 categories of secondary process cognition (e.g., abstraction, social behavior), 7 categories of emotions (e.g., positive affect, anxiety)</td>
<td>computer</td>
<td>instrumental thematic broad content</td>
</tr>
<tr>
<td>Analysis of Verbal Behavior</td>
<td>Weintraub (1981)</td>
<td>15 dimensions including pronouns (I, we, me), negatives (e.g., not, no, never), qualifiers (e.g., kind of), expressions of feelings (e.g., love), and adverbia l intensifiers (e.g., really, so)</td>
<td>judges</td>
<td>instrumental thematic broad style</td>
</tr>
<tr>
<td>TAS/C DICTION</td>
<td>Mergenthaler (1996)</td>
<td>Emotional tone, abstraction, referential activity</td>
<td>computer</td>
<td>instrumental thematic specific style</td>
</tr>
<tr>
<td></td>
<td>Hart (1984)</td>
<td>5 master variables (certainty, optimism, activity, realism, commonality) with a total of 35 linguistic subdimensions</td>
<td>computer</td>
<td>instrumental thematic broad style</td>
</tr>
<tr>
<td>LIWC</td>
<td>Pennebaker et al. (2001)</td>
<td>82 variables; standard linguistic dimensions (e.g., pronouns, articles), psychological processes (e.g., emotion words, causation words), relativity (past tense, inclusive words), personal concerns (e.g., school, religion, sexuality)</td>
<td>computer</td>
<td>instrumental thematic broad content/style</td>
</tr>
<tr>
<td>LSA</td>
<td>Landauer et al. (1998)</td>
<td>n/a; 2 strategies with focus either on low (content approach) or high (style approach) frequency words</td>
<td>computer</td>
<td>representational semantic n/a content/style</td>
</tr>
</tbody>
</table>
# Measure Latent Dimensions

## Communication Vagueness Dictionary

### Bluff and Recovery
- Actually (1)
- Alright (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)
- Everyone_knows (1)
- If_anything (1)
- If_nothing_else (1)
- In_any_case (1)
- In_any_event (1)
- In_a_way (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)
- Let's_see (1)
- Maybe (1)
- Not_sure (1)
- Sorry (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)
- More_or_less (1)
- Mostly (1)
- Much (1)
- Plenty_of (1)
- Quite_a_bit (1)
- Rather_large (1)
- Rather_small (1)
- Small_amount (1)
- Some (1)
- What_little (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)
- Normally (1)
- Presumably (1)
- Seems (1)
- Tendency (1)
Any Question?