Text Analytics, A Personal Retrospecitve

Presented to the 2015 FedCASIC Meeting by
James R. Caplan, PhD.

This presentation is my own and does not represent any Official Position of the Department of Defense

Text Analytics comes from Cognitive Psychology

- Cognitive psychology formed around how words and ideas are connected. Think Berkeley, Hume and John Stuart Mill from the 18th Century
- Reemerged in the 1950s based on the WWII focus on human performance and attention, developments in computer science, especially artificial intelligence, and interest in linguistics. Think Chomsky and McClelland
- Basis of my graduate training. 1968 Masters thesis based on word associations

Text Analytics = Concept Analysis

- Early studies were human-powered
- Subjects sorted statements into “buckets” based on how they seemed to “go together”
- Group would discuss results and reach consensus

Problems with Manual Concept Task

- Category definitions changed with ongoing context – requiring resorting
- Definitions were hard to keep in mind as number of buckets increased past 9 or 10
- Consensus-building arbitrary and results unreliable across groups

Text Analytics and Surveys

- Employee attitude surveys typically included open-ended questions, comments, and "Other/Specify" responses
- Contractors sanitized personal information, places, and expletives then categorized and coded answers

Problems with Survey Text

- Expensive, time-consuming
- Added months to final analysis, obviating the advantages of computer administration
- Eventually, open-ended questions were dropped from our employee surveys

What Did We Lose?

- Important way to know if some questions were confusing or ambiguous
- Lost alternatives we never considered
- The ability for respondents to interact with us: perhaps, an important positive motivator

Computer-assisted Coding

- SPSS comes up with “Text Analysis for Surveys,” approx. 2008
- Promise: automated coding and categorizing
- Reality: relies on data dictionaries and intensive human intervention- it’s a note taker
- Problem: extensive preparation required, no context sensitivity, no serious natural language processing
- Three versions later, no real improvement

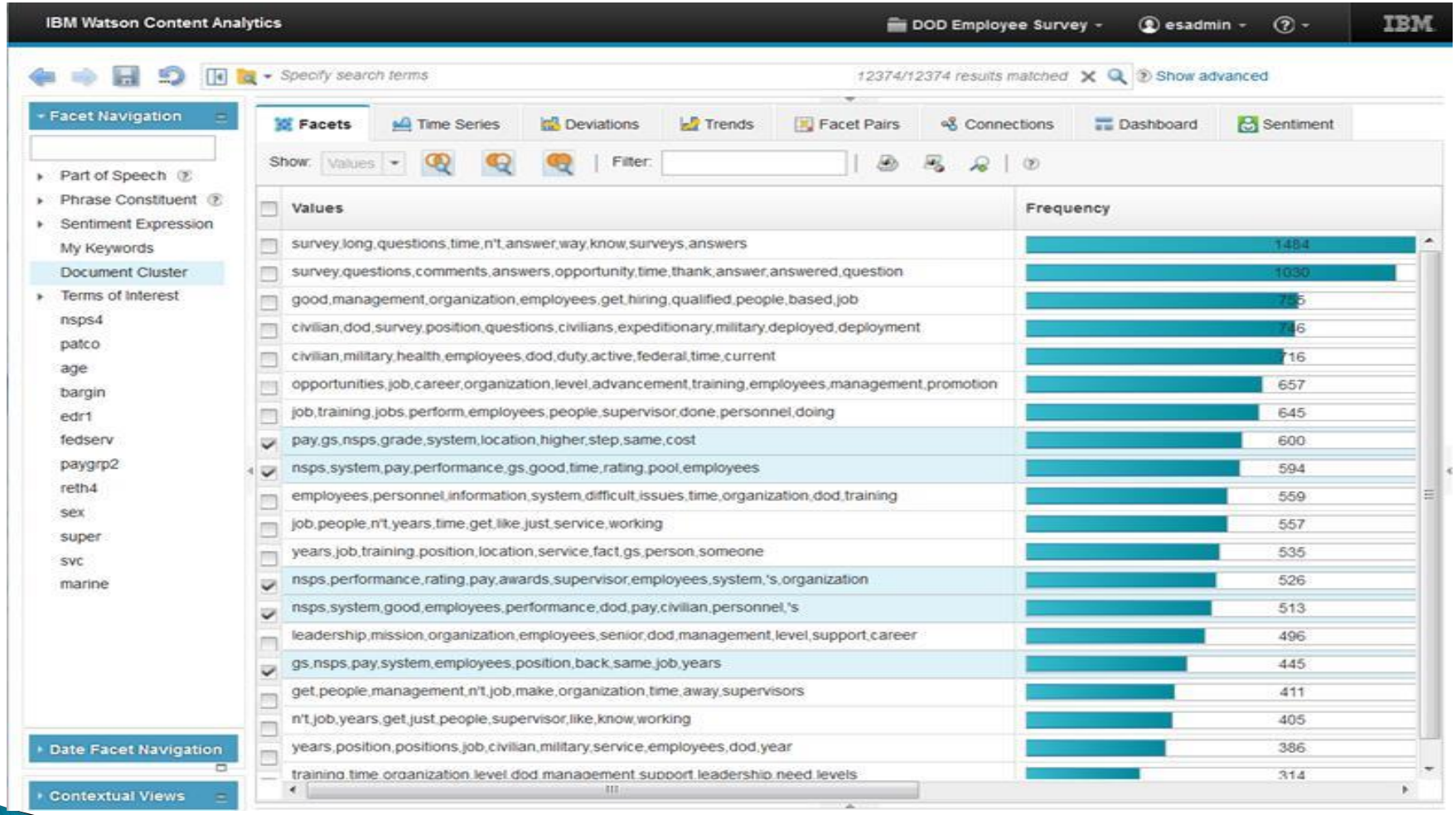
Further Advances

- 2009, IBM purchases SPSS
- Many other solutions emerge, Cognos, IBM Media Analytics, Watson, just by IBM
- Other solutions emerge but emphasis on marketing research, biological and medical research, brand and product preferences, analysis of Big Data, and national security

Where are We Now?

- Application of Watson to survey data
- Solves the problem of valence/affect (known as sentiment analysis by market researchers)
- Sanitized dataset from 2008
- Question: “If you have comments or concerns that you were not able to express in answering this survey, please enter them in the space provided. Any comments you make on this questionnaire will be kept confidential, and no follow-up action will be taken in response to any specifics reported.”
- Very preliminary results – Todd threw this into Watson with no instructions. I haven’t had a chance to interact with it yet

Initial Watson Clusters



Inspection of the Clusters

1. Comments about the Survey, itself (too long, redundant) (59.2% negative)
2. Comments about specific questions (50.2% negative)
3. Comments about the organization (57.7% negative)
4. Comments about work/job satisfaction (55.1% negative)
5. Etc.

Next Steps

- Try the Categorization function
- Add some key words to see if we can extract "obvious" clusters and identify non-obvious ones
- Refine our sentiment analysis
- See how clusters correlate with known demographics
- Check out some "Other/Specify" responses

Questions at the End of the Session, Please

james.r.caplan2.civ@mail.mil