Coding open answers

Peter Lugtig p.lugtig@uu.nl

Text data

- Ubiquitous in surveys
  - But rarely analyzed
- Even more so in big data
  - Social Media, speech (=text)

- What to do?
  - Historically: manually code them
  - Now/future: machine learning or combination

Coding open answers

Important to keep in mind how you will use the data.

- Are detailed codes necessary?
  - Population size, time, importance of question
- What was goal of the survey?
  - e.g. if it was to evaluate quality of teaching, focus on this.
- Are there existing coding schemes?
  - e.g. occupational, educational coding

How to code open answers manually

- Open
  - Break down, compare, and categorize data.
- Axial
  - Make connections between categories after open coding.
- Selective
  - Select the core category, relate it to other categories and confirm and explain those relationships.
  - See Boeije (2009) for more details

Example of coding of open answers

- "What is the most important problem facing our country?"

1. "abortion"
2. "economy, welfare should be checked into better, choices of how tax money is spent."
3. "honey, i really couldn't tell you."
4. "moral problem – general moral decay; drugs; crime; poor education; divorce; alcohol."
5. "our national debt"
6. "our physical responsible and the rest will fall into place"
7. "taxes, eats up your income"
8. "the federal bonus they keep promising to give to wwii veterans"
9. "the life amendment"

Example from ANES 1984-2000; Goodrich 2008

Example of coding of open answers (2)

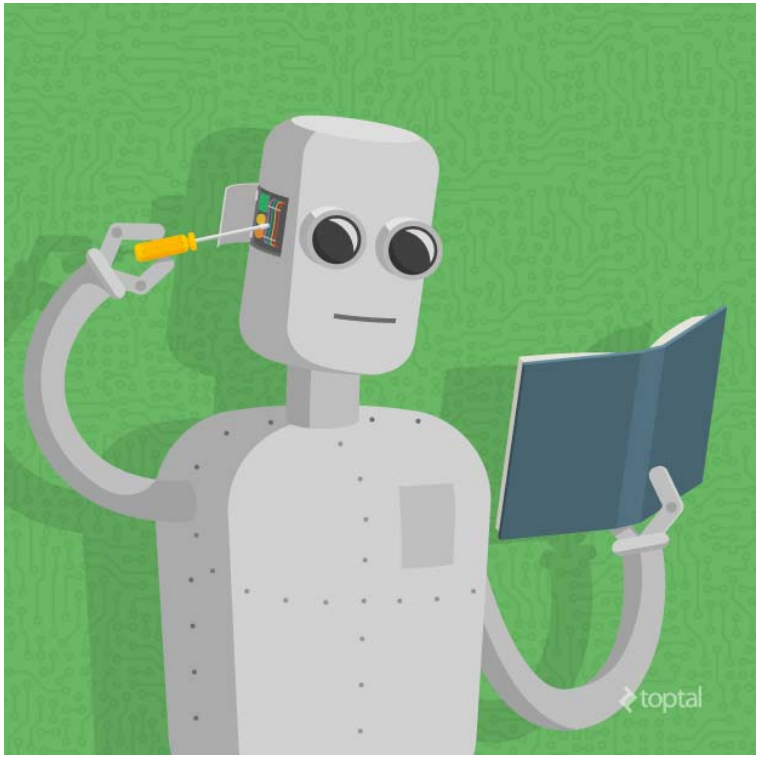
- Abortion
- AIDS/STDS
- **Arms/Weapons**
- Big Government
- **Budget**
- Business
- Campaign Finance Reform
- Central America
- **Children (not schools)**
- Communism
- Cost of Living
- **Crime (not drugs)**
- **Defense/Military**
- Discrimination/Race
- **Drugs**
- Drunk Driving
- **Economy**
- **Education**
- **Employment**
- Energy
- **Environment**
- Equality
- Families (not valu
- Far East
- Farming
- Guns
- **Healthcare**
- **Housing**
- **Hunger**
- Immigration
- Impeachment
- Imports/Exports
- Insurance
- International Relations
- Iran
- Iraq
- Japan
- Justice System
- Kuwait
- Media
- Medicare/Medicade
- **Middle East**
- **Morality**
- Native Americans
- Nicaragua
- **Nuclear**
- Oil/Gas
- Overpopulation
- Persian Gulf
- Pornography
- **Poverty**
- President
- Puerto Rico
- Religion
- **Russia**
- Saudi Arabia
- Seniors (not Social Security)
- Sexual Preference
- **Social Security**
- Somalia
- South America
- Space Program
- Star Wars
- **Taxes**
- Teen Pregnancy
- Terrorism
- Unions
- Wages
- **War/Peace**
- Wealth Inequality
- **Welfare**

Codes in bold: among 9 most frequently codes in any election between 1984 and 2000

Example from ANES 1984-2000; Goodrich 2008

Coding open answers automatically

- Large texts are impossible to code by hand
- Answer: Machine learning



Machine learning in coding text

- Large subfield in IT
  - Do counts of words
  - Sentiment analysis

Machine learning in coding text

- Sentiment analysis
  - Associations with topics
  - Using the twitter ‘hose’
  - U.S. election presidential Debate 2012

Barack Obama

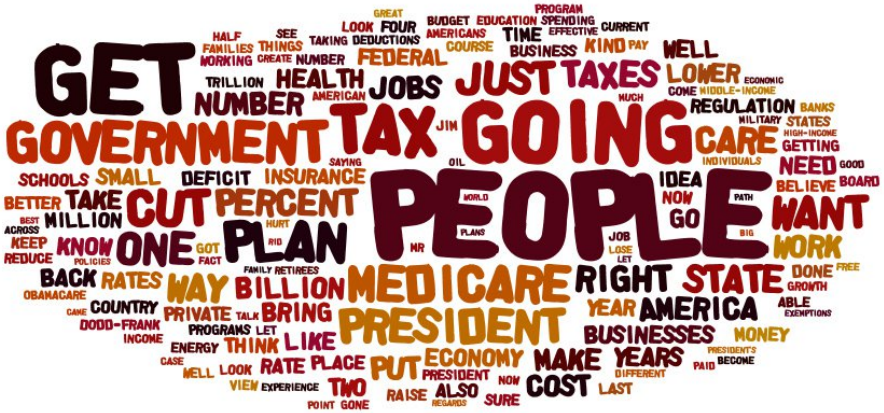
THE DAILY CALLER

Presidential debate, Oct. 3, 2012

Machine learning in coding text

- Sentiment analysis
  - Associations with topics
  - Using the twitter ‘hose’
  - U.S. election presidential Debate 2012

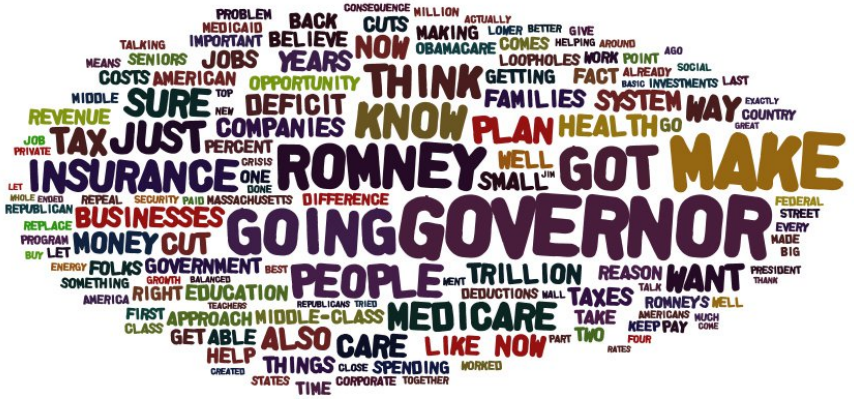
Mitt Romney



THE DAILY CALLER

Presidential debate, Oct. 3, 2012

Barack Obama

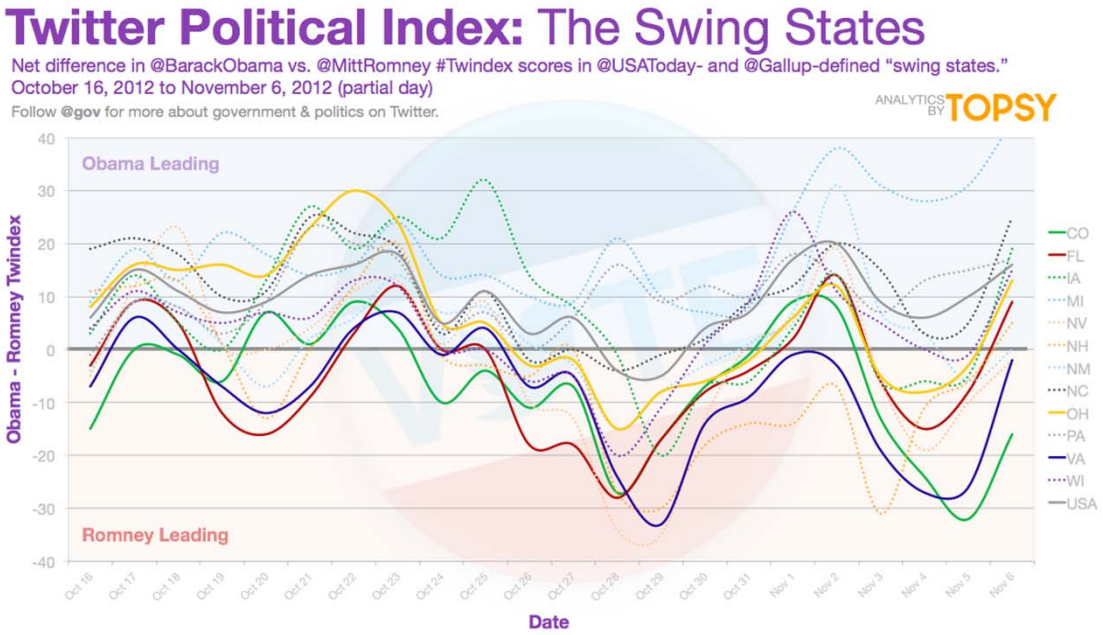


THE DAILY CALLER

Presidential debate, Oct. 3, 2012

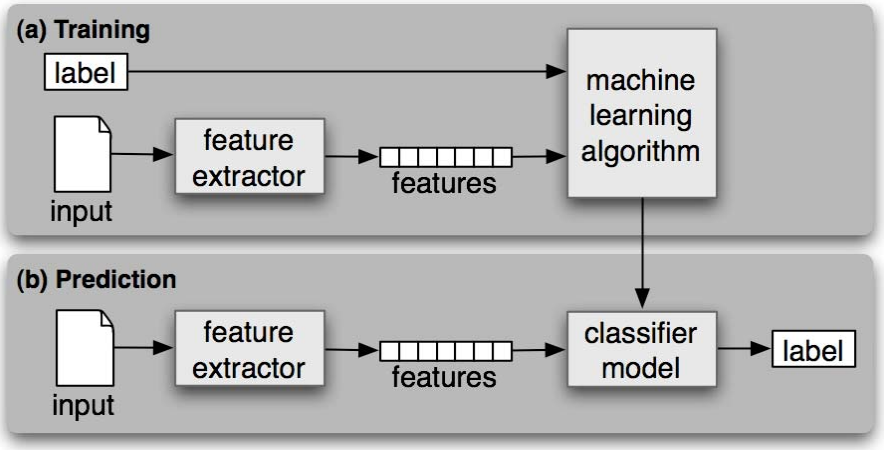
Machine learning in coding text

- Large subfield in IT
  - Do counts of words
  - Frequencies of text fragments -> group these
    - Sentiment analysis



Machine learning in coding text

- Sentiment analysis
  - How to do it
  - Supervised learning using a training set
  - Support Vector Machines
    - See Bird, Klein and Loper (2014)
    - Acccuracy with human coding: ~90%



Machine learning

- Sentiment analysis not very sophisticated
- Other approaches relevant for survey research
  - Structural Topic Models (Roberts et al 2012;2014)

Structural Topic Model

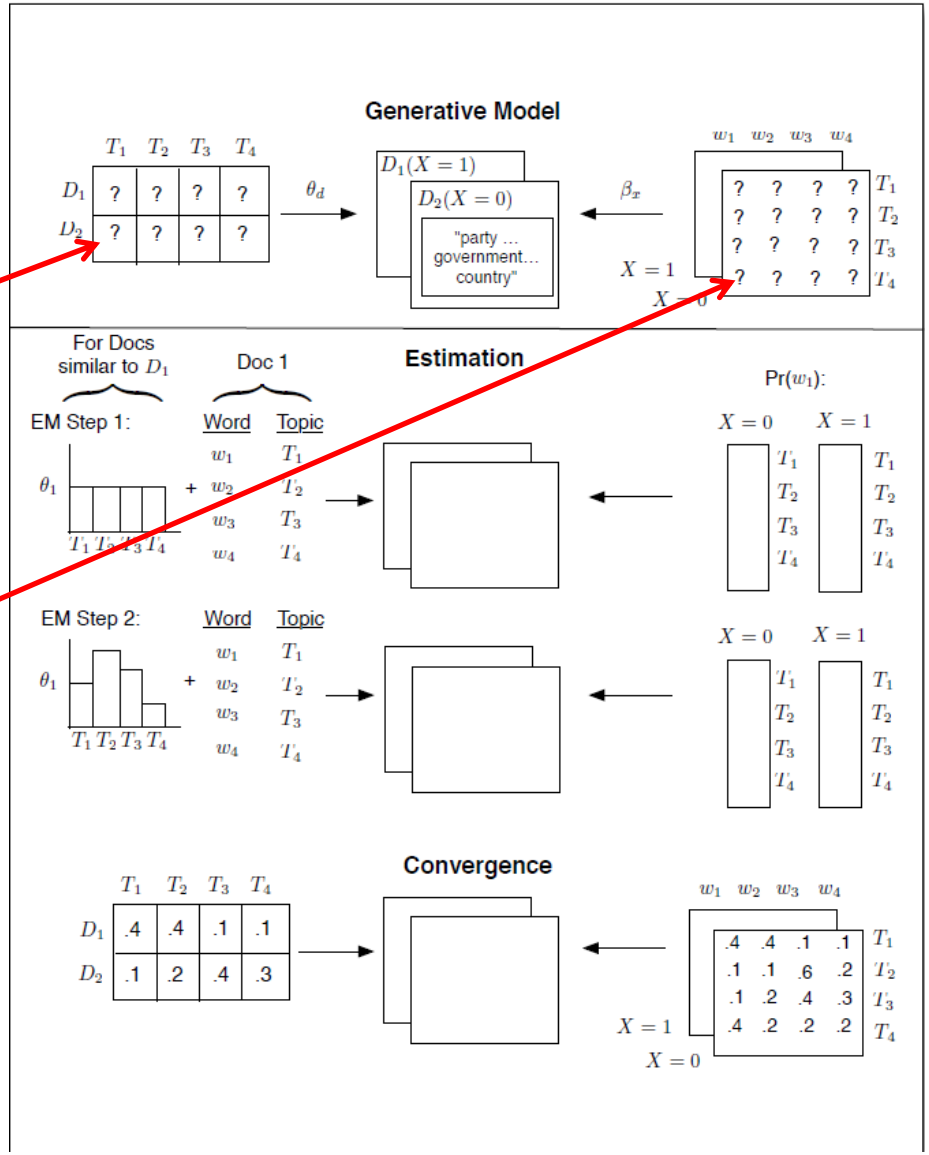
- Mixture models (as in statistics)
  - LDA (Latent Dirichlet Analysis)

- 3-level mixture model
  - 1. Document (case) is mixture of topics
  - 2. Topic is mixture of words
  - 3. words
- Between the levels a log-normal link function (multinomial)

Structural Topic Model

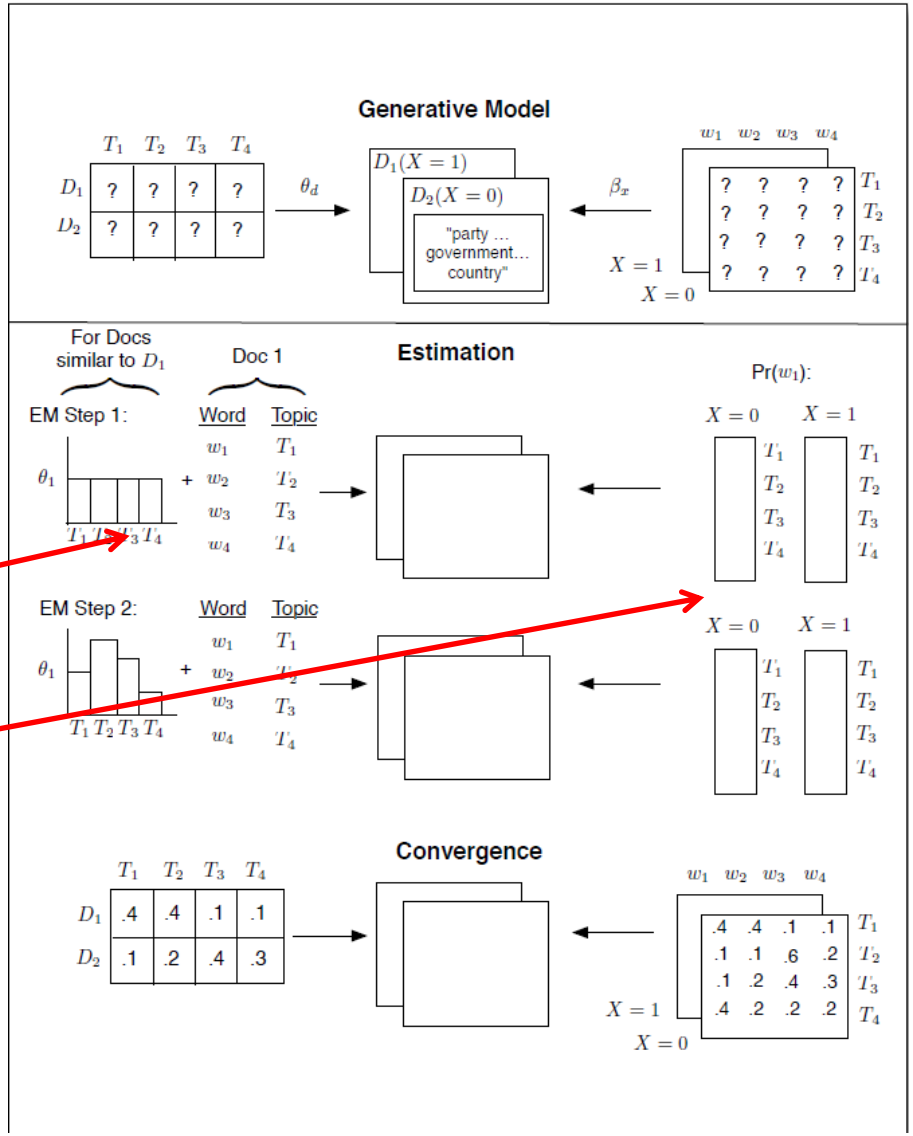
- Heuristic model
  - from Roberts et al 2016

- Generation:
- (Gibbs sampler)
- Multiple Topics (T) within documents(D)
- Multiple Words (w) within topics (T)



Structural Topic Model (2)

- Heuristic model
  - from Roberts et al 2016
- Estimation:
- Per document (case):
- Words and topics form document
- How likely is every word related to topic

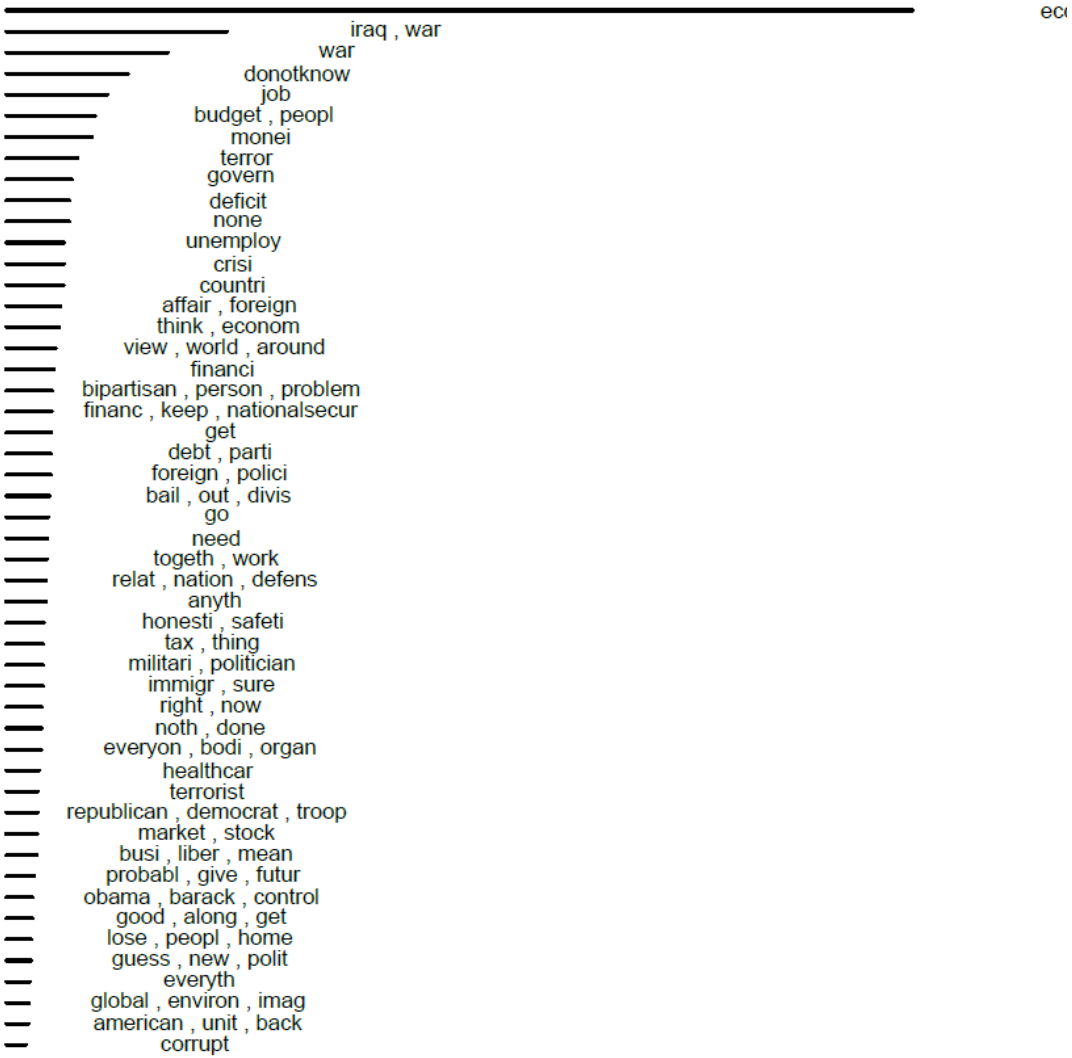


Structural Topic Model (3)

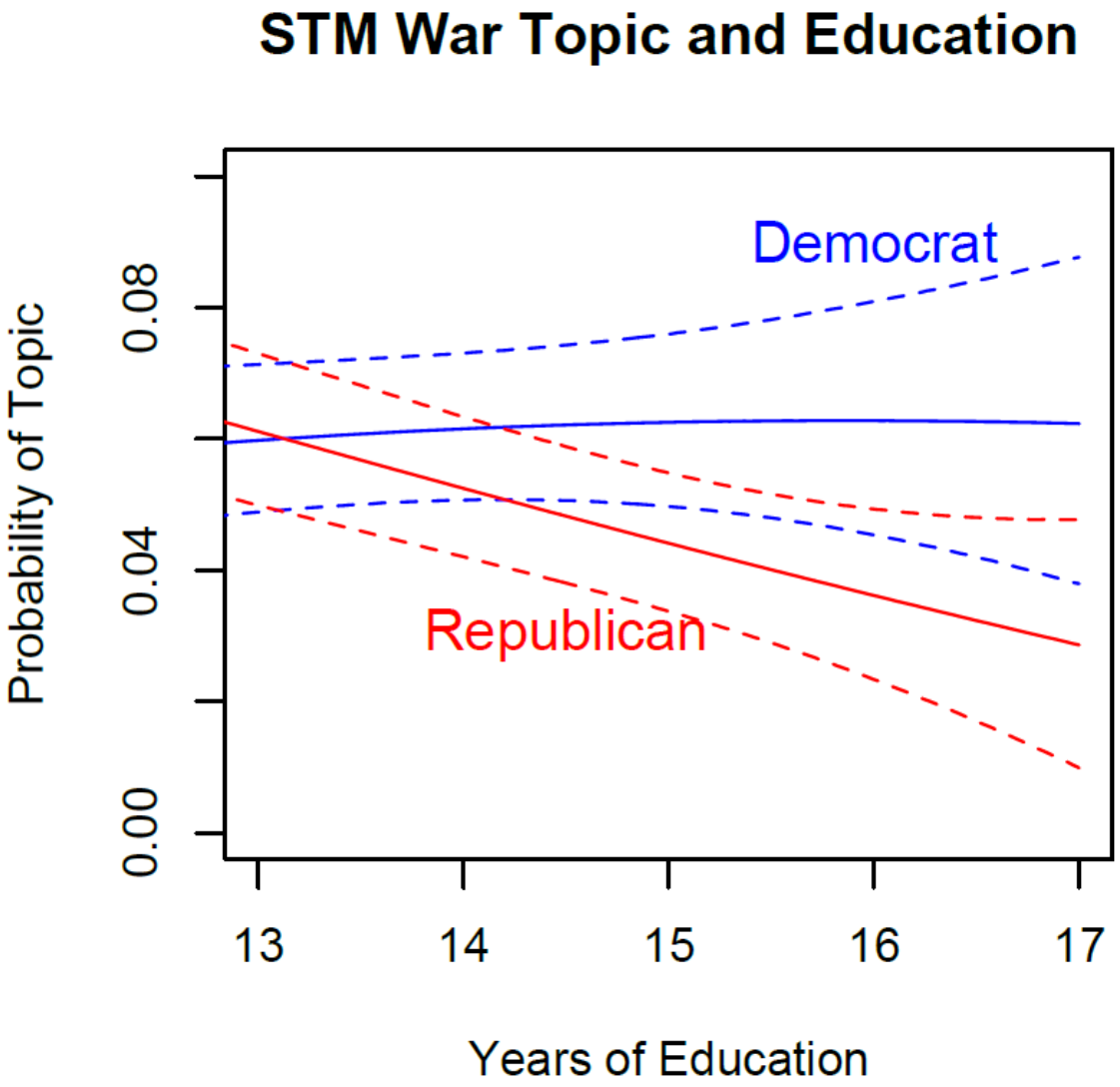
- Mixture models (as in statistics)
  - LDA (Latent Dirichlet Analysis)

- Can be linked to covariates
  - Do proper statistical analysis
  - E.g. do males use specific words or word-clusters more often?

The STM package in R - output



The STM package in R - output



Comparison with manual coding

- Abortion
- AIDS/STDS
- **Arms/Weapons**
- Big Government
- **Budget**
- Business
- Campaign Finance Reform
- Central America
- **Children (not schools)**
- Communism
- Cost of Living
- **Crime (not drugs)**
- **Defense/Military**
- Discrimination/Race
- **Drugs**
- Drunk Driving
- **Economy**
- **Education**
- **Employment**
- Energy
- **Environment**
- Equality
- Families (not values)
- Far East
- Farming
- Guns
- **Healthcare**
- **Housing**
- **Hunger**
- Immigration
- Impeachment
- Imports/Exports
- Insurance
- International Relations
- Iran
- Iraq
- Japan
- Justice System
- Kuwait
- Media
- Medicare/Medicade
- **Middle East**
- **Morality**
- Native Americans
- Nicaragua
- **Nuclear**
- Oil/Gas
- Overpopulation
- Persian Gulf
- Pornography
- **Poverty**
- President
- Puerto Rico
- Religion
- **Russia**
- Saudi Arabia
- Seniors (not Social Security)
- Sexual Preference
- **Social Security**
- Somalia
- South America
- Space Program
- Star Wars
- **Taxes**
- Teen Pregnancy
- Terrorism
- Unions
- Wages
- **War/Peace**
- Wealth Inequality
- **Welfare**

STM analysis: example

- Conclusions:
  - Manual
  - Supervised (using training data – not worked out today)
  - Unsupervised (STM approach)
- Manual and Automatic coding yield similar data
  - 94% of raw responses yield the same code (Roberts et al 2016)
  - In a lot less time
  - Automatic coding yields other benefits
    - Statistics on classification quality etc.
    - Statistical models

References

- Boeije (2009) Analysis in qualitative research. Sage: London.
- Bird, S., Klein, E. & Loper, E. (2014) the natural Language Processing with Python. http://www.nltk.org/book/
- Roberts, M.E., Stewart, B.D., Tingley, D., Lucas, C., Leder-Luis, J., Kushner gadarian, S., Albertson, B., Rand, D.G. (2014) Structural topic models for open ended survey responses. American Journal of Political Science, 58. 1064-1082. DOI:10.1111/ajps.12103
- Roberts, M.E., Stewart, B.M., & Tingley, D. (forthcoming) stm: R package for Structural Topic Models. Journal of Statistical Software
- Goodrich (2008) A coding methodology for open-ended survey questions. Paper for the 2008 new faces in political methodology conference.