POIR 613: Measurement Models and Statistical Computing

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Course website:

pablobarbera.com/POIR613/

Today

1. Solutions for last week’s challenge
2. Reminder: peer feedback was due yesterday
   - Please submit via Blackboard as well
3. Next deadline: October 30 for descriptive statistics
4. Other announcements:
   - No class on November 21st
   - Office hours at regular time tomorrow
5. Advanced topics in text analysis:
   - Event detection
   - Ideological scaling
   - Word embeddings

Overview of text as data methods

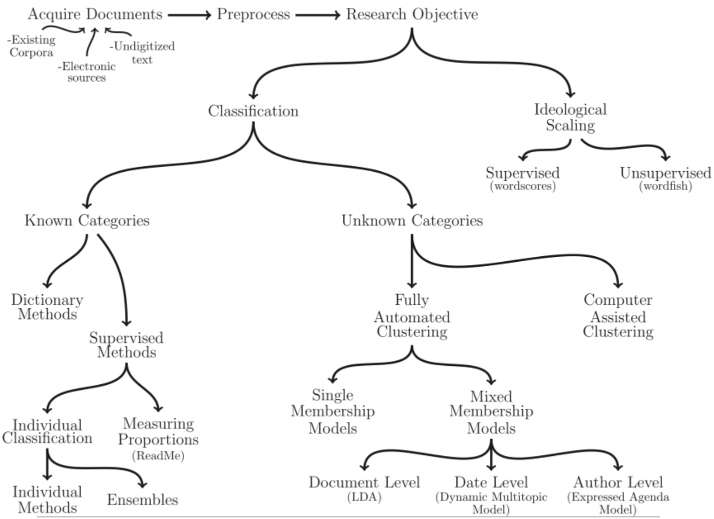

Fig. 1 in Grimmer and Stewart (2013)

Event detection in textual datasets

Event detection (Beieler et al, 2016)

Goal: identify who did what to whom based on newspaper or historical records.

Methods:

- Manual annotation: higher accuracy, but more labor and time intensive
- Machine-based methods: 70-80% accuracy, but scalable and zero marginal costs
  - Actor and verb dictionaries; e.g. TABARI and CAMEO.
  - Named entity recognition, e.g Stanford's NER

Issues:

- False positives, duplication, geolocation
- Focus on nation-states
- Reporting biases: focus on wealthy areas, media fatigue, negativity bias
- Mostly English-language methods

Ideological scaling using text as data

Wordscores (Laver, Benoit, Garry, 2003, APSR)

- Goal: estimate positions on a latent ideological scale
- Data = document-term matrix $\mathbf{W_R}$ for set of “reference” texts, each with known A_{rd}, a policy position on dimension d.
- Compute $\mathbf{F}$, where F_{rm} is relative frequency of word m over the total number of words in document r.
- Scores for individual words:
  - $P_{rm} = \frac{F_{rm}}{\sum_r F_{rm}} \rightarrow$ (Prob. we are reading r if we observe m)
  - Wordscore $S_{md} = \sum_r (P_{rm} \times A_{rd})$
- Scores for “virgin” texts:
  - $S_{vd} = \sum_w (F_{vm} \times S_{md}) \rightarrow$ (weighted average of scored words)
  - $S^*_{vd} = (S_{vd} - \overline{S_{vd}}) \left(\frac{SD_{rd}}{SD_{vd}}\right) + \overline{S_{vd}} \rightarrow$ Rescaled scores.

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician *i*, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \mathrm{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is “loquaciousness” of politician *i*
  ψ_m is frequency of word *m*
  β_m is discrimination parameter of word *m*

- Estimation using EM algorithm.
- Identification:
  - Unit variance restriction for θ_i
  - Choose *a* and *b* such that $\theta_a > \theta_b$

Word embeddings

[LINK]