Midterm Review

Format

- 4-6 Primary questions (with subparts)
- Understanding of when, how, and why to use methods we've talked about in class
- Many questions are open-ended with multiple possible responses
  - [1-2 sentence answers are typically sufficient, no need to write novels]
- Basic definitions

- If you need any equations we will give them to you (you should be able to recognize them but don't need to memorize them)

Word Statistics

- Key ideas behind two popular methods for examining word statistics:
  - Log-odds with a Dirichlet prior ("Fightin' Words")
  - Pointwise mutual information scores
- Examples of applications and understanding of when these methods are useful

Topic Models

- Motivation behind topic modeling
- High-level understanding of LDA formulation and inference
- Key strengths and limitations of LDA
  - When later variants of LDA might be more useful
- Agenda setting and framing
- Structured Topic Model (STM)
  - Example of LDA-variant designed for text analysis
  - Understanding of key differences from LDA and why they are useful

Word Embeddings: Construction

- Intuitive ideas behind representing words as vectors
- Distributional Hypothesis
- TF-IDF weighting
- Word2Vec
  - Difference between CBOW and Skip-gram
  - Skip-gram model and training
  - Practical challenges (e.g. negative sampling)

Word Embeddings: Applications

- Example applications:
  - Measuring bias (gender bias / occupational stereotypes)
  - Measuring change in word meanings over time
  - Measuring stereotypes over time
- Embedding manipulation:
  - Cosine similarity, Euclidean distance
  - Gender subspace
  - Averaging keywords
- Evaluations:
  - Analogy tasks, similarity benchmarks, extrinsic metrics
  - Comparisons with hand-curated analyses or benchmarks
  - Comparisons with survey or crowd-sourced data

Affect and Lexicons

- Emotions:
  - Different models of emotions in psychology
- Lexicons:
  - When lexicons are useful and when they are not
  - Different variants of lexicons
    - Categorical vs. continuous, directed (connotation frames) vs. not
    - Examples: LIWC, NRC lexicons, connotation frames
  - Lexicon construction
    - Manual vs. automated
- Data annotating:
  - Likert scale, Best-worst scaling

Data Annotation

- Tips and tricks for components of annotation process
  - Data selection, annotator selection, task design, quality control
  - Examples: Decomposition, context and priming, quiz questions
- Annotator agreement metrics
  - Percent Agreement, Cohen’ s Kappa, Fleiss’ Kappa, Krippendorff’s alpha
- Ethics of crowdsourcing
- Qualitative coding [from Adam Koon’s lecture]
  - General ideas and procedure
  - Deductive vs. inductive coding

Classification Models

- Logistic Regression
- Neural networks
- Prevalence Estimation
  - Classify and count, adjusted classify and count, probabilistic classify and count

Hypothesis Testing

- Basic idea behind hypothesis testing
- Ability to interpret use of hypothesis tests by others
- Ability to select a hypothesis test based on data characteristics
- Selection bias, Simpson's paradox

Causal Inference

- Potential outcomes notation
- Definitions of core concepts:
  - ITE, ATE, Confounder, Mediator, Collider, observational study, randomized control trial
- Assumptions of causal inference:
  - Ignorability/exchangeability, conditional exchangeability / unconfoundedness, positivity, no interference, consistency
- Adjustments:
  - Regression, matching, propensity scores, stratification, IPW

Causal Inference with text

- Motivations
  - Example applications, challenges of working with text
- Adjustment Methods
  - Topic Inverse Regression Matching
  - Causal text embeddings

Network Metrics

- Motivations and examples where methods are useful
- Basic metrics:
  - Network density, closeness centrality, quarter-power scaling
- ERGM
- Graph Neural Networks