Exploratory text analysis for computational social science

CORE Congress
10 December 2020
Ian Stewart, University of Michigan

Group introductions (if time)

Introduction

- **Goal of tutorial**: understand and implement basic exploratory text analysis techniques for social science applications.

- **Background**: introductory statistics/probability, basic Python
  - Includes walk-through of code notebooks with time for exploration.

- **Information**: where to access data/code?
  - **Save time by navigating to notebooks now!**
    https://github.com/ianbstewart/CORE_tutorial_2020

    cutt.ly/CORE_tutorial

  - Should work in Firefox, Chrome.

Social science

- How do people make **social decisions** in everyday life? What explains the **structure** of conversations, relationships, communities?

- “Umbrella” for sub-disciplines:
  - Sociology
  - Social psychology
  - Geography
  - Communications
  - Political science
  - Linguistics
  - Social computing

Social science research lifecycle

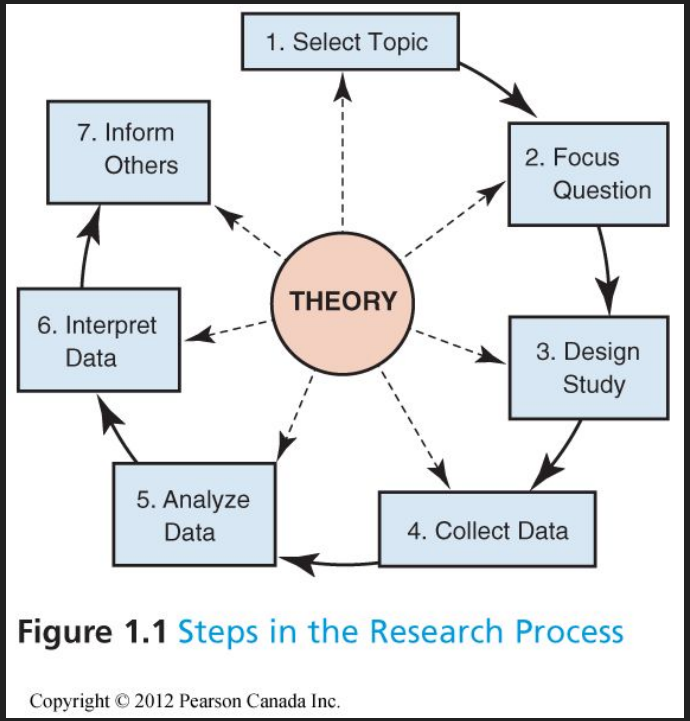
- Unlike many ML projects, social science research doesn't always start with a specific **task** or objective to "solve."

- Big goals are often **testing theories** of socialization, which requires scientific process: hypothesis, procedure, data collection, analysis, conclusion...

Social science research lifecycle

- Unlike many ML projects, social science research doesn't always start with a specific **task** or objective to "solve."

- Big goals are often **testing theories** of socialization, which requires scientific process: hypothesis, procedure, data collection, analysis, conclusion...



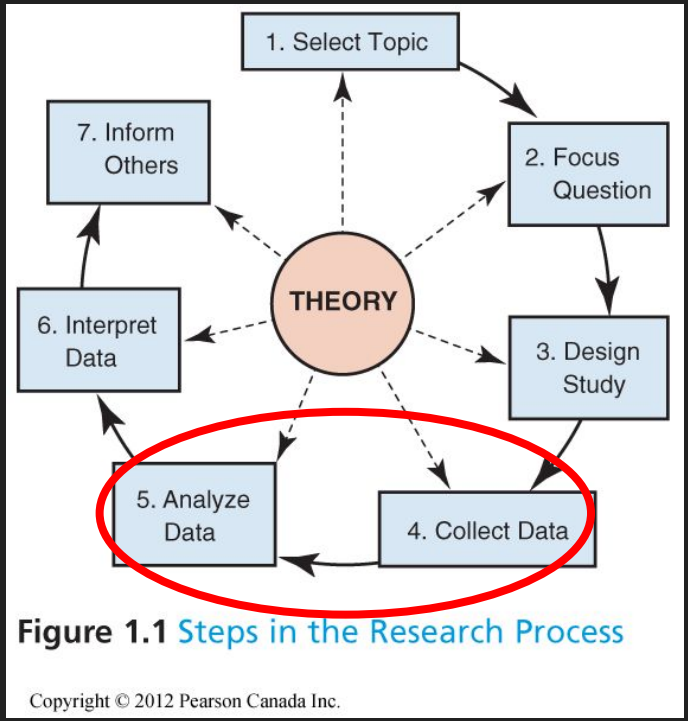
Babbie (2012)

Social science research lifecycle

- Unlike many ML projects, social science research doesn't always start with a specific **task** or objective to "solve."

- Big goals are often **testing theories** of socialization, which requires scientific process: hypothesis, procedure, data collection, analysis, conclusion...

- We focus on exploration because it's **creative** and relies on a variety of methods to "**think through**" data.



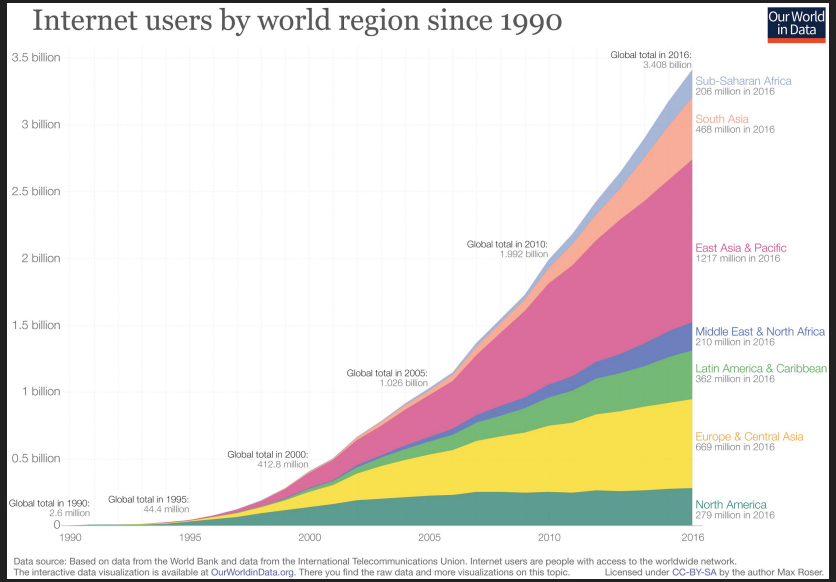
Babbie (2012)

Computational social science

- Traditional social science techniques: useful for building theory; not scalable to large/changing environments.

- CSS reveals insight from growing **digital trace data** through text and network analysis.

- Focus on underlying large-scale **social dynamics**: homophily, information spread, opinion shifts.



World Bank (2016)

Computational social science examples

Everyone's an Influencer: Quantifying Influence on Twitter

Eytan Bakshy*
University of Michigan, USA
ebakshy@umich.edu

Jake M. Hofman
Yahoo! Research, NY, USA
hofman@yahoo-inc.com

Winter A. Mason
Yahoo! Research, NY, USA
winteram@yahoo-inc.com

Duncan J. Watts
Yahoo! Research, NY, USA
djw@yahoo-inc.com

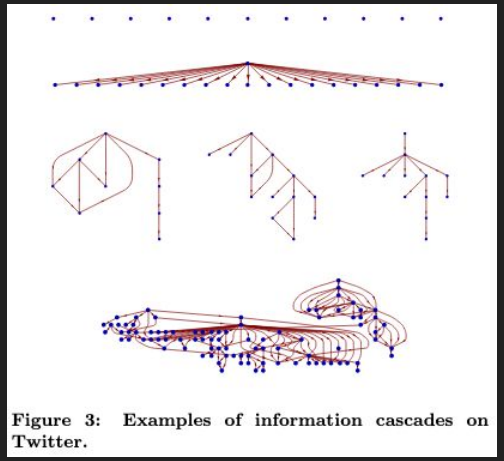
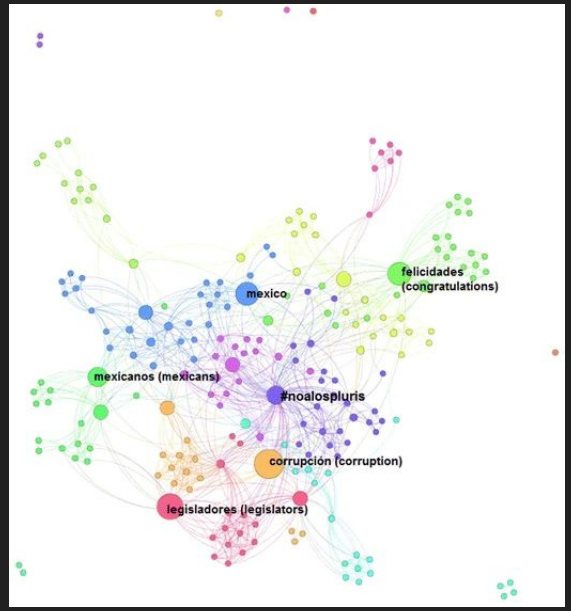


Figure 3: Examples of information cascades on Twitter.

Detecting sociosemantic communities by applying social network analysis in tweets

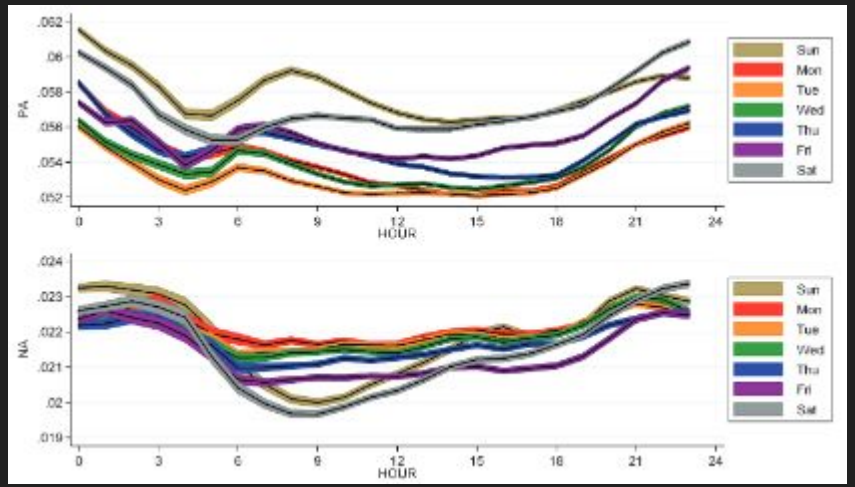
Rocío Abascal-Mena ✉, Rose Lema & Florence Sèdes



Computational social science examples

Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures

Scott A. Golder* and Michael W. Macy



Language from police body camera footage shows racial disparities in officer respect

Rob Voigt[a,1], Nicholas P. Camp[b], Vinodkumar Prabhakaran[c], William L. Hamilton[c], Rebecca C. Hetey[b], Camilla M. Griffiths[b], David Jurgens[c], Dan Jurafsky[a,c], and Jennifer L. Eberhardt[b,1]

Example	Respect Score
First Name, Ask For Agency, Questions: [name], can I see that driver's license again? It- it's showing suspended. Is that- that's you? (Disfluency, Negative Word, Disfluency)	-1.07
Informal Title, Ask For Agency, Adverbial "Just": All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick. ("Hands On The Wheel")	-0.51
Apology, Introduction, Last Name: Sorry to stop you. My name's Officer [name] with the Police Department.	0.84

Interdisciplinary benefits

Computer science	Social science
Study anything	Study social things
Methods driven	Question driven
Large found data	Small designed data
Prediction	Explanation

(Wallach 2015)

Text analysis (NLP)

- Typical goals: help computers automate complicated language tasks like information **retrieval**, **recommendation**, and **dialogue**.

- For social scientists: text analysis can help extract communication patterns from large-scale data, such as groups of words that characterize documents (topics).

Text analysis example

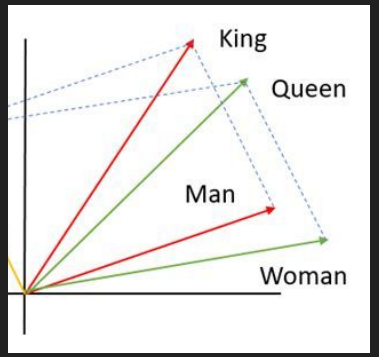


Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

Garg et al. (2018)

Domain: fake news

- Misinformation about current events is rampant online, particularly in spaces with limited moderation practices (Gillespie 2018).

- Many platforms are interested in **detecting** fake news before it can spread (Shu et al. 2017), which requires a strong understanding of the news content.

- What aspects of **text content and style** differentiate fake news from real news?

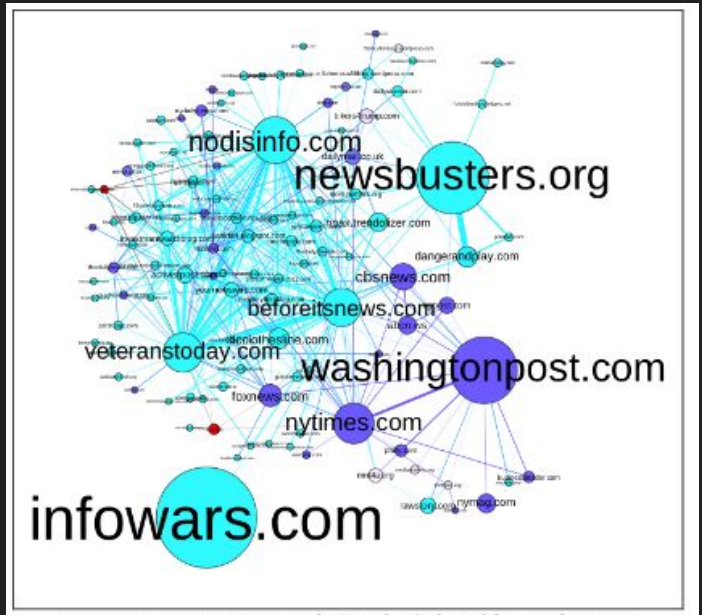


Figure 1. Domain Network Graph, Colored by Media Type
Purple = mainstream media; Aqua = alternative media;
Red = government controlled media

Starbird (2017)

Domain: fake news data

- For this tutorial we will explore two data sets related to fake news.

1. Crowdsourced fake news rewritten from real news stories (Pérez-Rosas et al. 2018).

Domain: fake news data

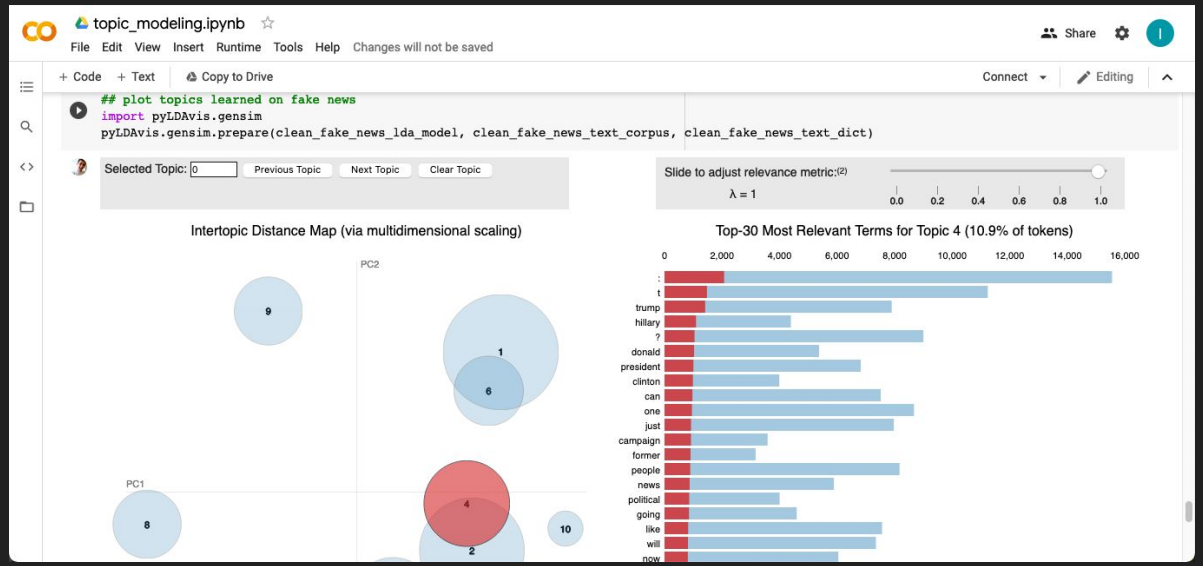
- For this tutorial we will explore two data sets related to fake news.

1. Crowdsourced fake news rewritten from real news stories (Pérez-Rosas et al. 2018).

2. Fake news and real news articles from U.S. media in 2017,
   e.g. Reuters and CNN (“real”); AddictingInfo and 100PercentFedUp (“fake”).
   a. Caveat: data collection strategy unclear.
   b. Full data here: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Tutorial format: Colab

- Google Colaboratory hosts interactive Python code “notebooks”.
- We’ll use several notebooks to explore fake news data.



Outline

- Introduction
- Word frequency
- Topic modeling
- Word embeddings
- Wrap-up

Word frequency

Word occurrence

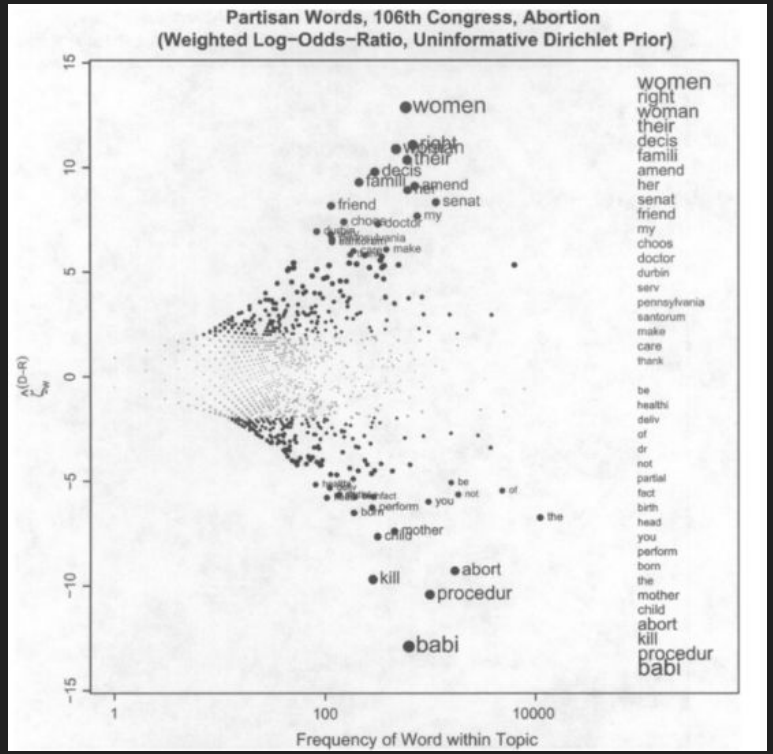
- Counting individual words/phrases can reveal clear patterns in data to probe further.
  - In political discussions: “shooting” implies focus on violence.

- It can also reveal surprising patterns!
  - “Mother,” “daughter” can reflect “family” framing of women’s issues.

Word occurrence: example study

- To understand **framing** in U.S. politics, Monroe et al. (2008) compared word frequency in political speeches from different parties.

- When discussing abortion, Republican politicians focused more on children (“**child**”, “**baby**”) whlie Democratic politicians focused on women’s rights (“**women**”, “**decision**”).



Monroe et al. (2008)

Word frequency

- First step in exploration: what words occur **most frequently** in the data set?
- Easy solution: count all words, sort by frequency.
- Problems?

!".	851
249	384
!	343
100th	278
'	182
20th	161
24	113
3bn	90
2010	89
allows	86
650m	76

Top-10 words in crowdsourced fake news (after filtering)

Word frequency

- First step in exploration: what words occur **most frequently** in the data set?

- Easy solution: count all words, sort by frequency.

- Problems?

    - Function words (“the”, “a”) are most frequent but also **least informative**.

    - Documents have **variable lengths**! Longer documents contribute more words.

    - The **interesting** words (e.g. “conspiracy”) may occur only a few times, but occur more frequently in fake news versus real news.

Normalized frequency

- Solution 1: convert to relative frequency for better representation.

- Use Maximum Likelihood Estimation to compute probability of observing a word.

P(word) = P(observe word in randomly chosen document)

P(word) = #(word) / sum_(all words)#(word)

TF-IDF

- Solution 2: normalize term frequency by document frequency.

- Intuition: we want to identify words that characterize a few documents (higher **information value**), rather than words that occur across all documents.

TF-IDF(word) = #(word) / #(documents containing word)

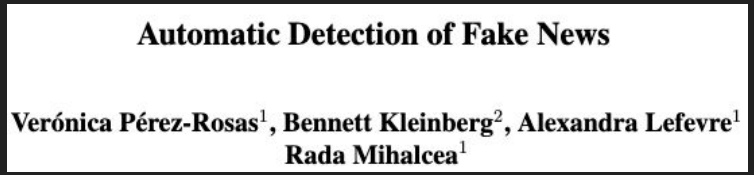
Frequency ratio

- What if we want to know the words that occur more often in fake news than real news?

- Solution 3: compute ratio of word occurrence.
  - Edge case: how to deal with 0 counts?

$$\text{frequency\_ratio}(word) = P(word, data\_1) / P(word, data\_2)$$

Case study: fake news


Automatic Detection of Fake News

Verónica Pérez-Rosas[1], Bennett Kleinberg[2], Alexandra Lefevre[1]
Rada Mihalcea[1]

- Let's look at the fake news data generated by crowdsourced workers in Pérez-Rosas et al. (2018).

- **Key question**: what specific words are most frequent in fake news articles?

- Open word_frequency.ipynb in Colab

Additional exploration results

What did you find?

Possible extensions

- Smoothing word frequency can fix 0-counts and help find rare words.

- **Stemming** and **lemmatization** can identify similar words for easier counting: e.g. “swimming”, “swims” reduce to “swim”.

- **Named Entity Recognition** can identify larger word units to count: e.g. personal names (“Donald Trump”) and locations (“San Francisco”).

Topic modeling

Document representations

- With dense representation of documents (bag-of-words), it's often difficult to find sub-groups of words that consistently characterize documents.

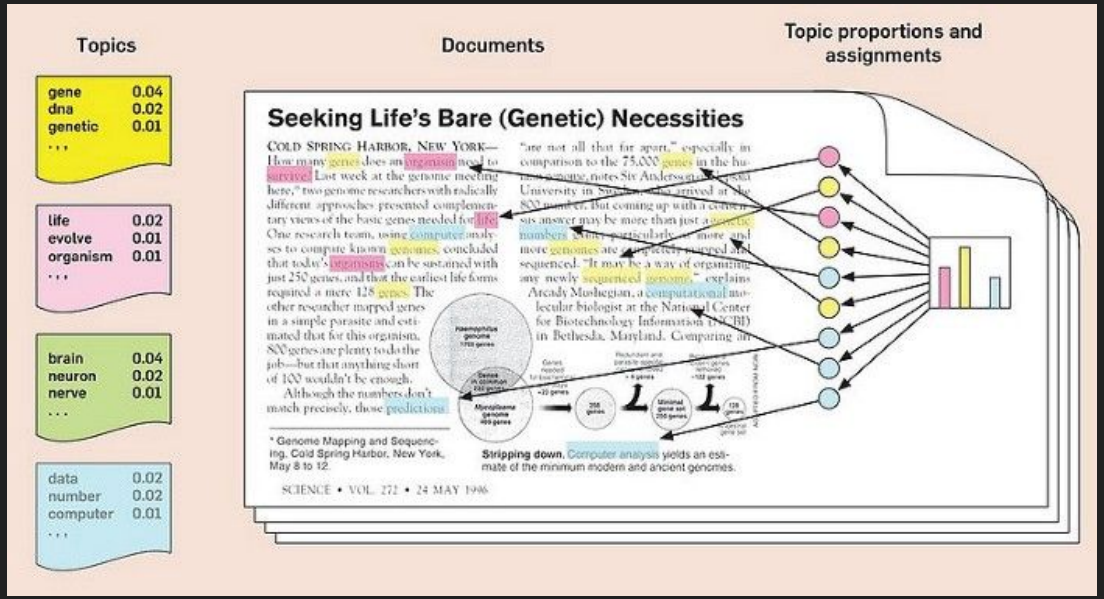
- How can we simplify the representation to capture more **general information** about the document?

- **Goal**: represent each document with latent "topics" that group similar words together.

Topic model overview

- Topic models try to “fit” latent dimensions to the observed text data.

- Models include both non-generative (no parameters) and generative versions (parameters, assumptions about word distributions).



Topic model: example

- To understand **polarization**, Demszky et al. (2019) analyzed the topics discussed in reactions to shootings on social media.

- Authors often discussed topics related to **news sharing**, **solidarity**, and **policy suggestions**.

Topic	10 Nearest Stems
news (19%)	break, custodi, #breakingnew, #updat, confirm, fatal, multipl, updat, unconfirm, sever
investigation (9%)	suspect, arrest, alleg, apprehend, custodi, charg, accus, prosecutor, #break, ap
shooter's identity & ideology (11%)	extremist, radic, racist, ideolog, label, rhetor, wing, blm, islamist, christian
victims & location (4%)	bar, thousand, california, calif, among, los, southern, veteran, angel, via
laws & policy (14%)	sensibl, regul, requir, access, abid, #gunreformnow, legisl, argument, allow, #guncontolnow
solidarity (13%)	affect, senseless, ach, heart, heartbroken, sadden, faculti, pray, #prayer, deepest
remembrance (6%)	honor, memori, tuesday, candlelight, flown, vigil, gather, observ, honour, capitol

Possible approaches

- What are some possible methods to identify groups of words that characterize discrete topics in documents?

Singular Value Decomposition

- Singular Value Decomposition converts a single matrix (X) into separate components, then reconstructs the matrix using its k-largest singular values.

$$X = U\ S\ V^T$$

$$S^* = S(\text{top-k columns}),\ V^* = V(\text{top-k rows})$$

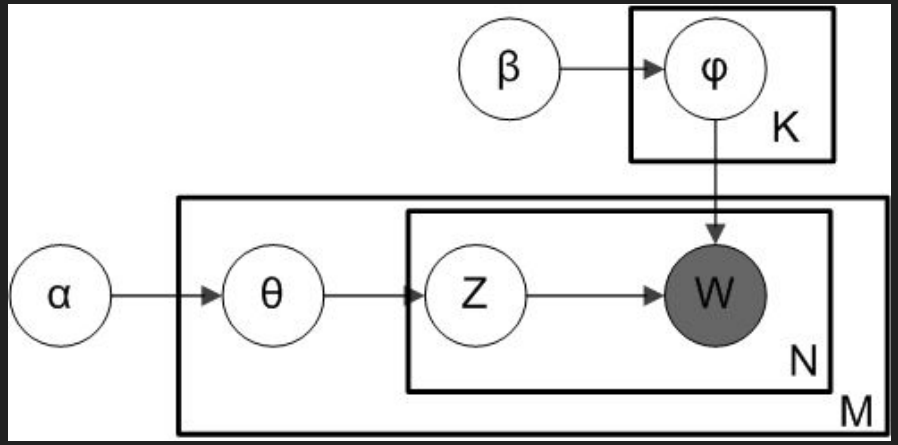
$$X^* = U\ S^*\ V^{*T}$$

- This identifies words that covary and therefore represent latent **topics** mentioned in the documents.

Latent Dirichlet Allocation

- Assume that document is composed of discrete **latent topics** that “generate” the words observed.

- Requires iterative sampling to fit observed data to underlying distributions.



$$
\begin{aligned}
\boldsymbol{\varphi}_{k=1\ldots K} &\sim \text{Dirichlet}_V(\boldsymbol{\beta}) \\
\boldsymbol{\theta}_{d=1\ldots M} &\sim \text{Dirichlet}_K(\boldsymbol{\alpha}) \\
z_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_K(\boldsymbol{\theta}_d) \\
w_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_V(\boldsymbol{\varphi}_{z_{dw}})
\end{aligned}
$$

Choosing the “best” model

- Even with just SVD and LDA, it’s not always clear which model is best suited for a given corpus.

- We can measure the utility of a topic model based on:

    - **Coherence**: do the words within each generated topic actually tend to occur in similar contexts in the data?

    - **Overlap**: do the words tend to overlap across topics?

    - **Perplexity**: can the topics trained from one corpus explain the distribution of words in a held-out corpus with high probability?

Case study: fake news

- We'll now try topic modeling on a larger dataset of fake news and real news articles from the "wild."

- **Key question**: what topics do fake news articles consistently cover, and are these topics really that different from real news?

- Open topic_modeling.ipynb in Colab

Additional exploration results

What did you find?

Possible extensions

- **Hierarchical** topic models can identify “layers” of topics to further explore discourse within single topic.

- **Guided** topic models learn to build topics around “seed” words, based on prior domain knowledge.

- **Author-document** topic models learn priors for writers: useful for social media (many “documents” per author).

Abortion		**Security/War**	
Join	Prolife	Killed	Military
Religious	Killed	Syrian	Illegal
Stand	Born	Military	Russian
Support	Unborn	Fast	Targeting
Conservative	Aborted	Furious	Back
Gun laws		**Immigration**	
Illegal	Dont	Join	Top
Free	Free	Support	Enter
Dont	Stop	Back	Check
Vote	Illegal	Stand	Stop
Stop	Give	Proud	Join

Joshi et al. (2016)

Semantic representations

Distributional semantics

- “You shall know a word by the company it keeps” (Firth 1957).

- Assumption: words that occur in similar contexts must have similar denotational meaning.

- “The man went to the store.”
  “The woman went to the store.”
  =>
  “man” has similar semantic value as “woman”

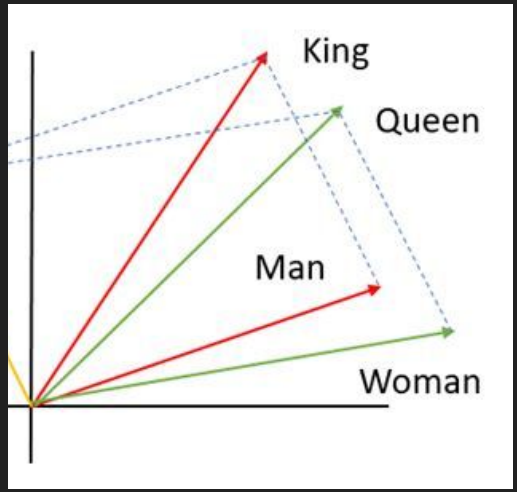
Word embeddings

- Common approach: train an unsupervised model to **maximize similarity** in representation space between words that occur in similar contexts.

- The semantic value of each word is represented with a dense “embedding” vector.

- Embeddings have useful geometric properties!

  “king” - “queen” ~ “man” - “woman”
  => binary “gender” dimension



Semantics: example

- To understand latent **social attitudes**, Garg et al. (2019) generated word embeddings from English news articles.

- For **racial** and **gendered** words, the authors found consistent cases of stereotypes via word connotations.

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

Garg et al. (2019)

Possible approaches

- If you have information about word co-occurrence, how might you learn embeddings for different words?

Glove

- Global Vectors for word representation (Pennington et al. 2014)

- Compute matrix of **global** word-word co-occurrence probability (within window), add bias for each word, down-weight frequent co-occurrences.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

- Minimize loss function for reduced-dimension embedding matrix (W) via gradient descent.

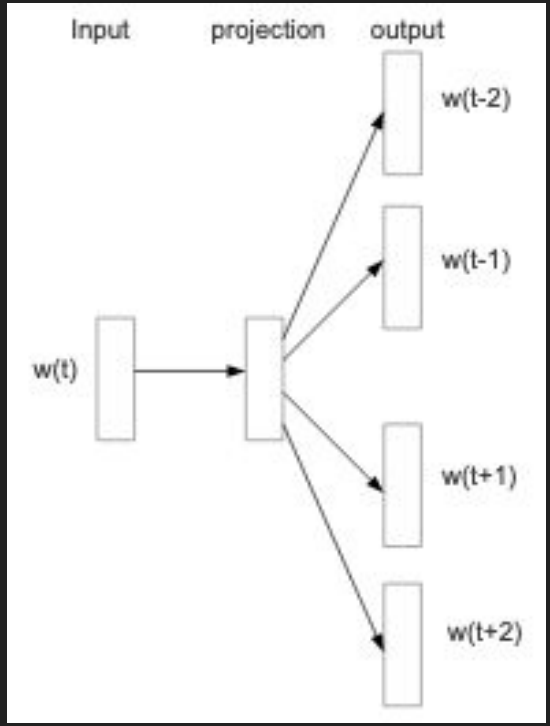
word2vec

- Continuous skip-gram with negative sampling (Mikolov et al. 2013)

- Train small neural network to **predict** the words that occur in the same “window” of a target word.

- Maximize similarity between target word and observed context, and minimize similarity between target word and negative sample contexts.

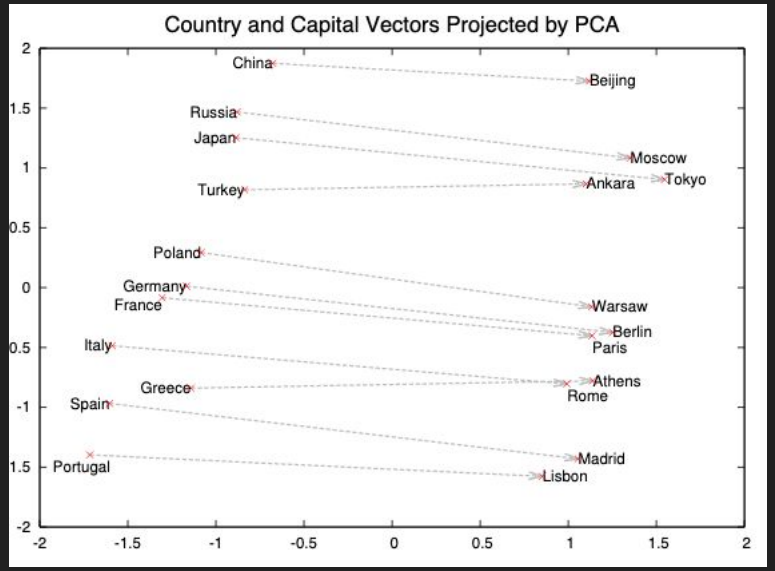
  Ex. Maximize sim(“cat”, “dog”), minimize sim(“cat”, “vehicle”).



Mikolov et al. (2013)

Word embedding: evaluation

- Word embeddings should encode **known word relationships**.
- Do embeddings encode **analogies** between words?
- Do embeddings encode **composition** of meaning?
- Note: same task may have same results for all datasets: analogies with “America” in tweets vs. news text; comparing “North” vs. “South”.



German + airlines	Russian + river	French + actress
airline Lufthansa	Moscow	Juliette Binoche
carrier Lufthansa	Volga River	Vanessa Paradis
flag carrier Lufthansa	upriver	Charlotte Gainsbourg
Lufthansa	Russia	Cecile De

Mikolov et al. (2013)

Case study: fake news discussions

- We'll train word embeddings on the same fake news and real news datasets that we investigated before.

- **Key question**: how are underlying attitudes toward particular words or phrases (e.g. "president") encoded in fake news articles?

- Open word_embeddings.ipynb in Colab

Additional exploration results

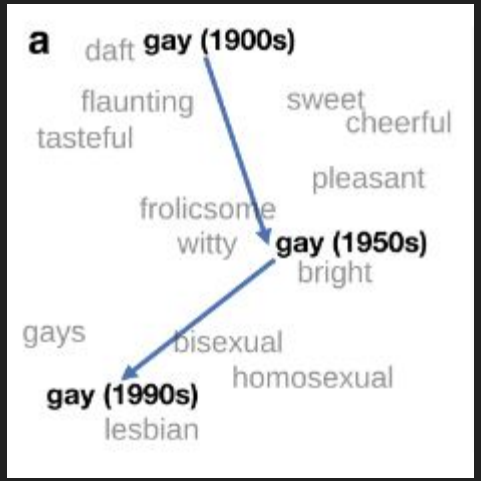
What did you find?

Possible extensions

- **Aligned** word embeddings between datasets (e.g. different time periods) can be directly compared.

- **Contextual** word embeddings capture the meaning of a word in its sentence:
  “I got money from the **bank**” vs.
  “I swam near the river **bank**”

- Testing for differences across sets of words (e.g. “woman”, “housewife”) can identify cases of **bias**.



Hamilton et al. (2016)

Wrap-up

- Areas covered
  - Word frequency
  - Topic models
  - Word embeddings

- By exploring fake news data, what have we learned?

Wrap-up

- Areas covered
  - Word frequency
  - Topic models
  - Word embeddings

- By exploring fake news data, we have learned:
  - Fake news focuses less on mainstream topics (finance, international relations) and more on **polarizing, “alternative”** perspectives (election fraud, presidential power).

  - Fake news frames the discussion of news topics to **encourage controversy**, e.g. highlighting identity politics of “People” and using “corruption” with respect to legal investigations.

  - This data set may be too “Trump” centric, based on the word associations that the models learned.

Choosing methods

- Choice of analysis method requires understanding study **goals**!

“Which method seems cool and exciting?” < “Which method helps to answer the research question better than standard approaches?”

Comparing methods: fill in the blanks

Method	Advantages	Disadvantages
Word frequency		
Topic models		
Word embeddings		

Comparing methods

Method	Advantages	Disadvantages
Word frequency	Easy to compute, clear interpretation, can compare across corpora, no parameters to define.	No context for word use, sensitive to irregular data, harder to understand high-level patterns.
Topic models		
Word embeddings		

Comparing methods

Method	Advantages	Disadvantages
Word frequency	Easy to compute, clear interpretation, can compare across corpora, no parameters to define.	No context for word use, sensitive to irregular data, harder to understand high-level patterns.
Topic models	Identifies word groups, scalable runtime/memory, shows document-level patterns, unsupervised methods have less inductive bias.	Need to define parameters, may get “tricked” into learning uninformative topics, sensitive to preprocessing, no word-level insight.
Word embeddings		

Comparing methods

Method	Advantages	Disadvantages
Word frequency	Easy to compute, clear interpretation, can compare across corpora, no parameters to define.	No context for word use, sensitive to irregular data, harder to understand high-level patterns.
Topic models	Identifies word groups, scalable runtime/memory, shows document-level patterns, unsupervised methods have less inductive bias.	Need to define parameters, may get “tricked” into learning uninformative topics, sensitive to preprocessing, no word-level insight.
Word embeddings	Captures context of word’s use, relationships between words, subtle connotations of usage.	Need to define parameters, hard to compare across models, longer runtime/memory, sometimes hard to interpret.

What do you think?

- Let's brainstorm! https://jamboard.google.com/d/1_kYUhcfeknPIHpIH_Qj5EDyfyhEI2sEUKKcSHf3ovlA/

- Which computational social science questions are you interested in studying with text analysis?

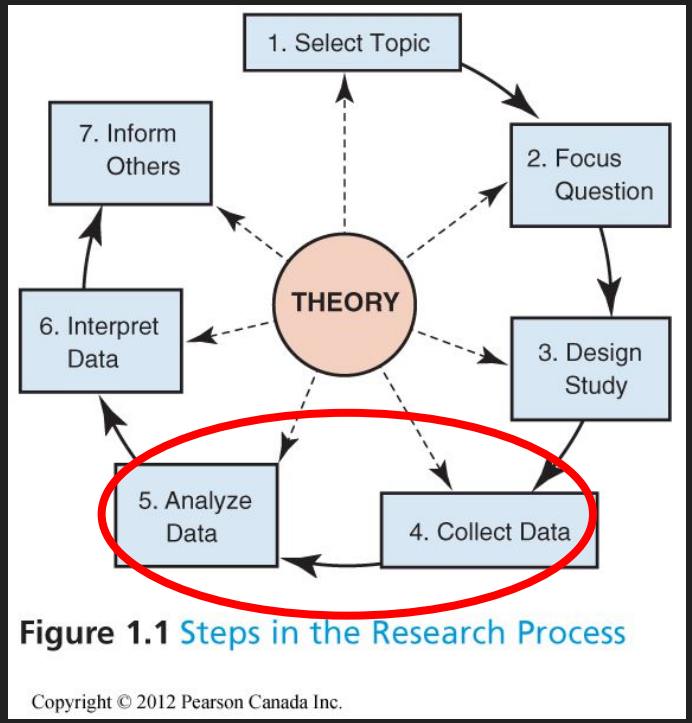
- How can the methods shown here help you explore your data more effectively?

- What would you have to **modify** about the methods to apply them to your situation?

Social science research lifecycle

- Data exploration is just one part of the cycle!

- Results of exploration may lead to:
  - More data collection, cleaning
  - Designing experiment to augment data
  - Shift toward supervised objective (e.g. fake news prediction)
  - Re-calibrating research questions
  - Adopting different theory



Babbie (2012)

Questions?

Thanks for participating!

Follow-up questions/thoughts ⇒ ianbstew@umich.edu