Advanced Topics & Ethics

Advanced Topics (topic modeling) + Ethics of big data

IMT 547 - Social Media Data Mining and Analysis

4-Mar-2021 (Week 9, Day 18)

Today's Topics

- Topic modeling
- Lab on topic modeling
- Reading for today: Ethics & limitations of big data
- Project work: Limitations of your project
- Next Class - office hours

Topic Modeling

Input: A document-term matrix (word order does not matter). Each topic will consist of a set of words where order doesn't matter, so we are going to start with the bag of word format.

Gensim: gensim is a python toolkit built for topic modeling. Popular topic modeling technique used LDA (Latent Dirichlet Allocation)

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are the topics in this set of document?**

Documents are the probability distribution or mix of these topics

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are these topics?**

Topic A: 40% banana, 30% kale, 10% breakfast…

What would you call topic A? Topic B?

Topic B: 30% kitten, 20% puppy, 10% frog, 5% cute…

Topics are a probability distribution or mix of words

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are these topics?**

Topic A: 40% banana, 30% kale, 10% breakfast…

FOOD

Topic B: 30% kitten, 20% puppy, 10% frog, 5% cute…

ANIMALS

Topics are a probability distribution or mix of words

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are the topics in this set of document?**

100% Topic A
FOOD

100% Topic B
ANIMALS

100% Topic B
ANIMALS

100% Topic A
FOOD

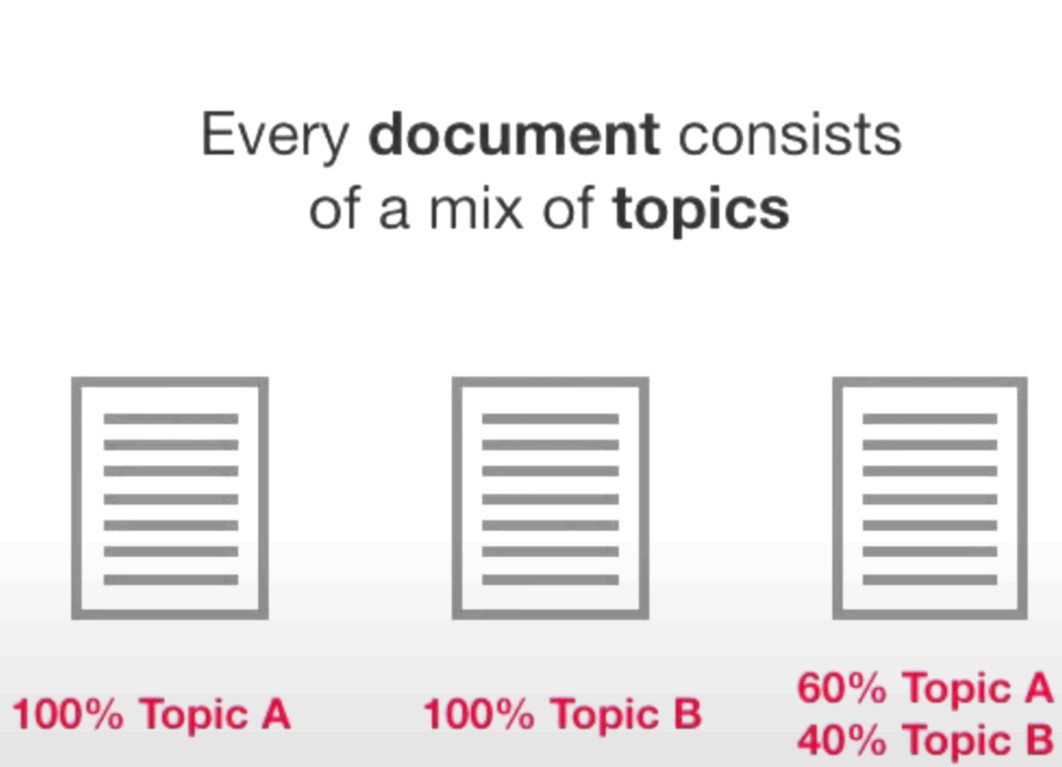
60% Topic A
40% Topic B

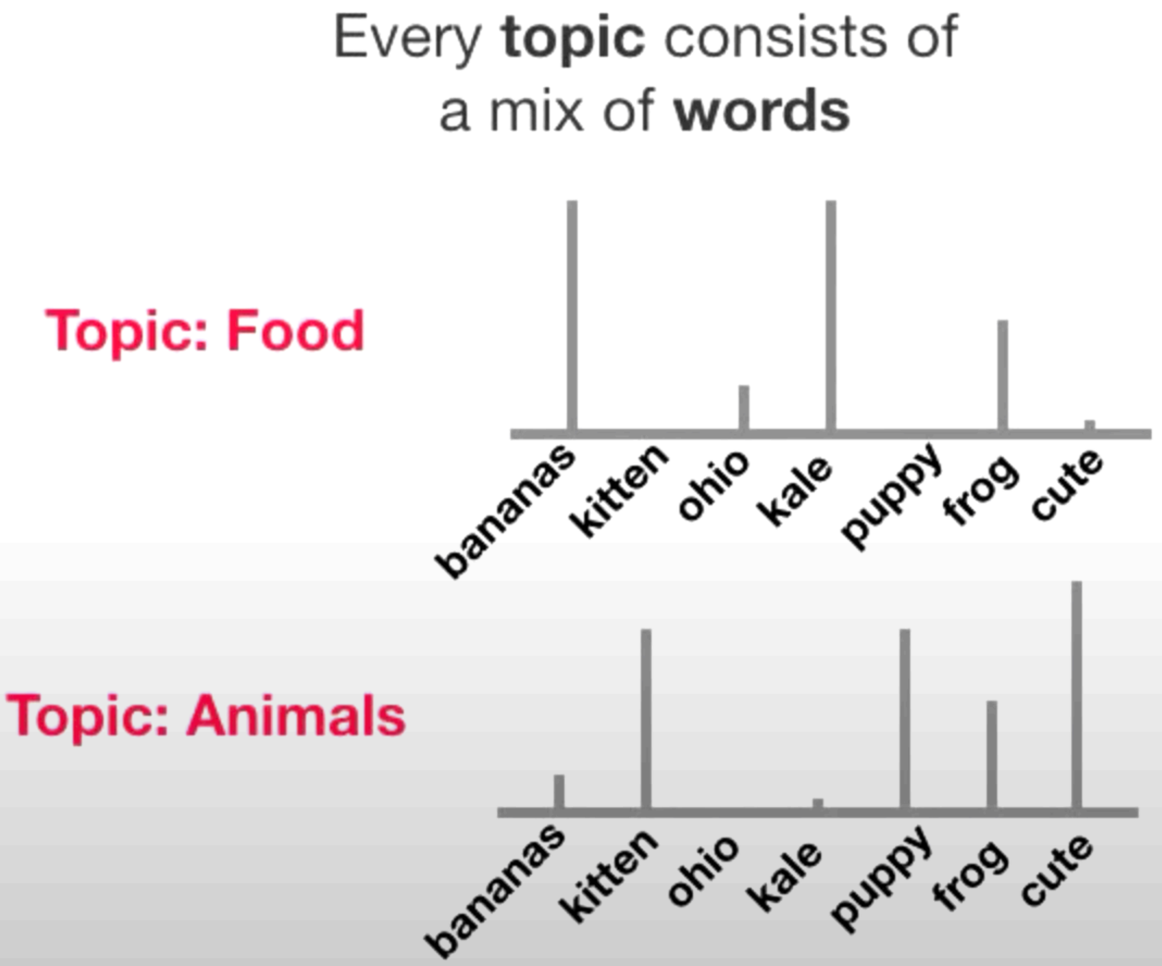
FOOD + ANIMALS

Documents are the probability distribution or mix of these topics

Topic modeling - LDA

Visualize the topic-word distributions





Topic modeling - LDA

How it works? (At a high level)

- Goal: To lean the topic mix in each document, and the word mix in each topic
- Choose the number of topics you think there are in your corpus. E.g.: K = 2
- Randomly assign each word in each document to one of 2 topics. E.g.: The word “banana” in Document # 1 is randomly assigned to topic B
- Go through every word and its topic assignment in each document. Look at (1) how often the topic occurs in the document and (2) how often the words in the topic overall. Based on this info, assign the word a new topic.
- Go through multiple iterations. Eventually the topics will start making sense. Interpret them

GENSIM takes care of these steps, especially the most complex steps.

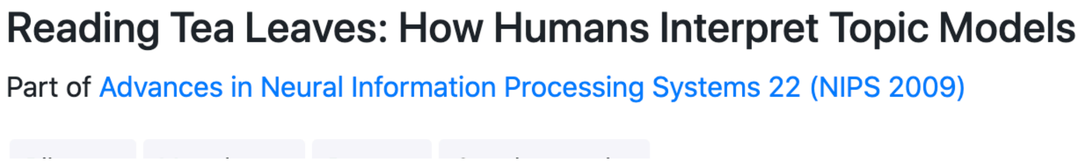
Topic Modeling

Input:

- A document-term matrix
- Number of topics
- Number of iterations

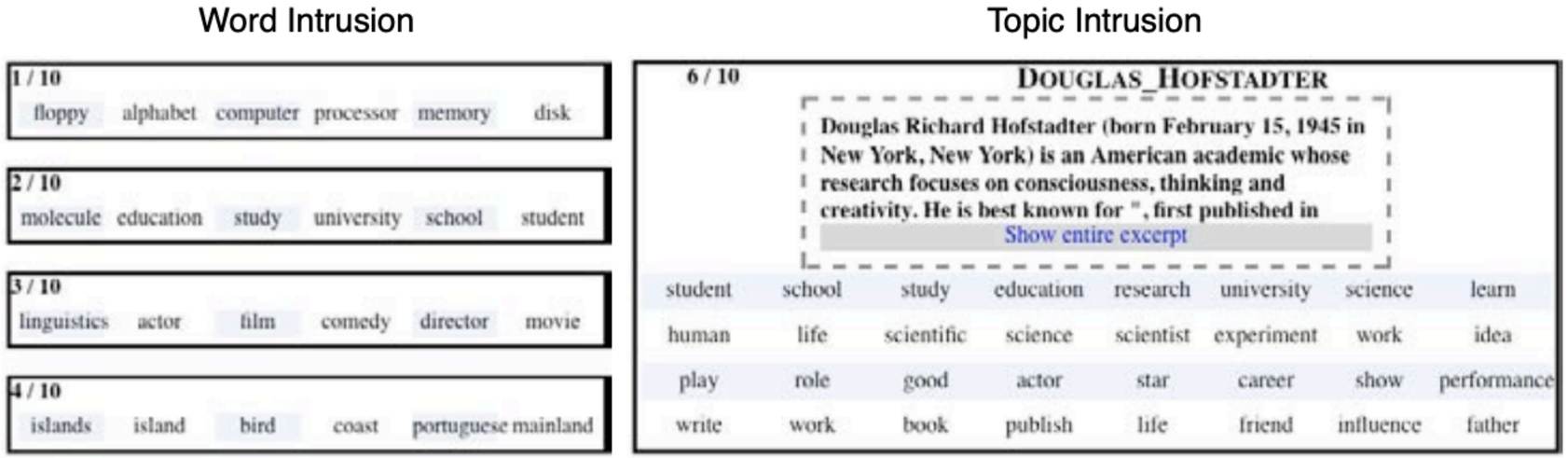
Gensim: gensim will go through the process to find the best word distribution for each topic and the best topic distribution for each document.

Output: The top words in each topic. Then human interpretation to figure out do these make sense or not. *Reading tea leaves!!*



Interpreting topic models

Paper: Reading Tea Leaves: How Humans Interpret Topic Models



Two human tasks. In the **word intrusion task** (left), subjects are presented with a set of words and asked to select the word which does not belong with the others.
In the **topic intrusion task (right)**, users are given a document's title and the first few sentences of the document. The users must select which of the four groups of words does not belong.

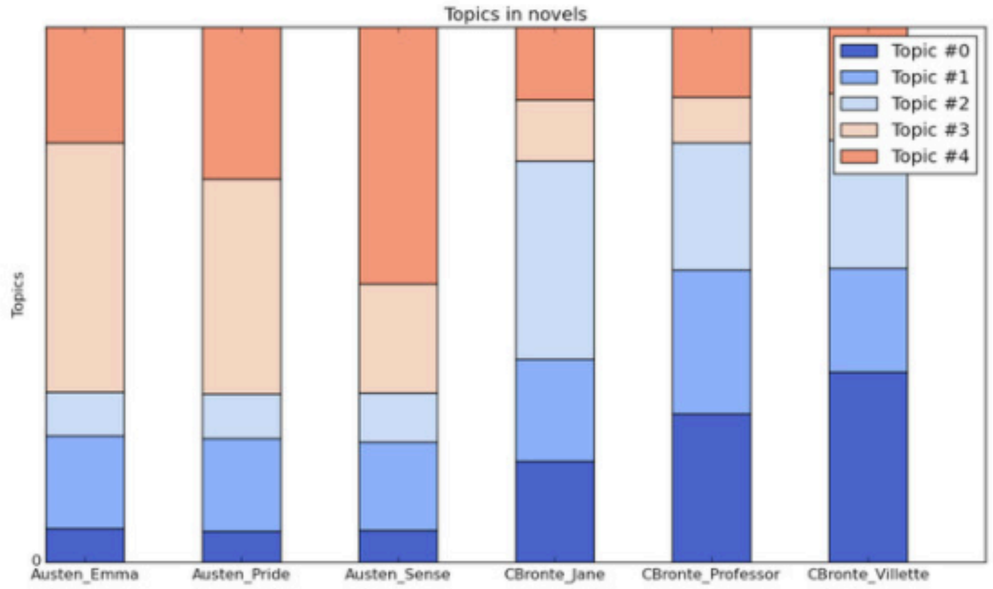
Interpreting topic models

Key questions that you would want to ask

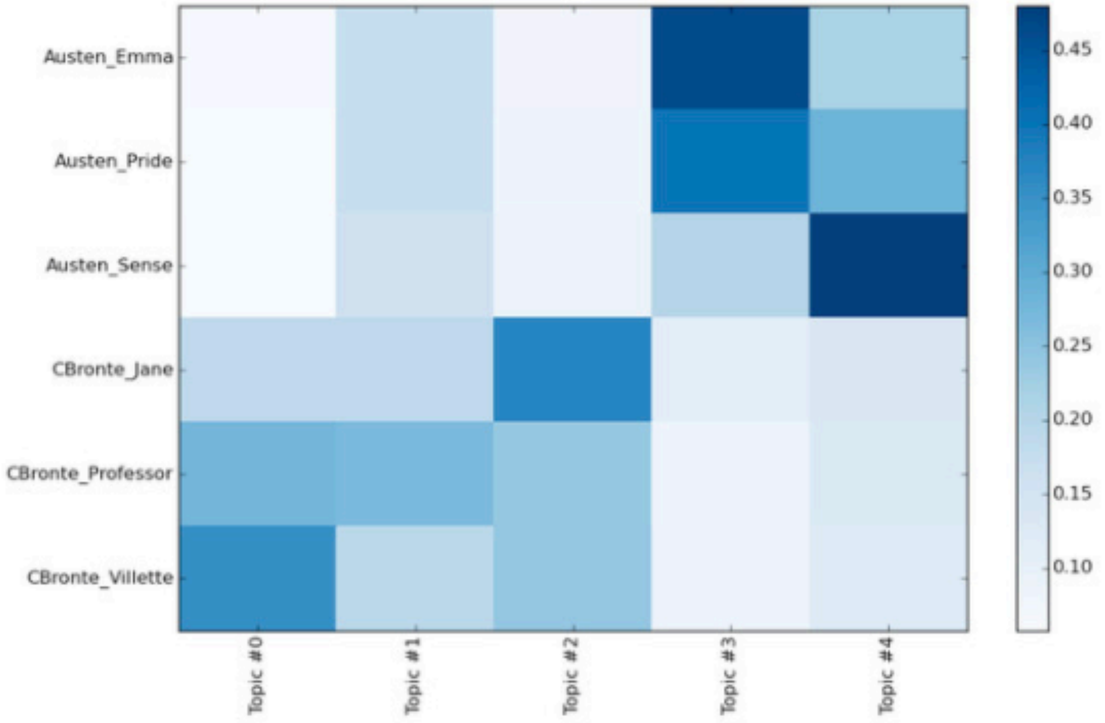
- What is the meaning of each topic?
- How prevalent is each topic?
- How do the topics relate to each other?
- How do the documents relate to each other?

Interpreting topic models via visualizations. ***What would be some good visualizations?***

Interpreting topic models via visualization



Stacked bar charts



Heat maps

Interpreting topic models via visualization

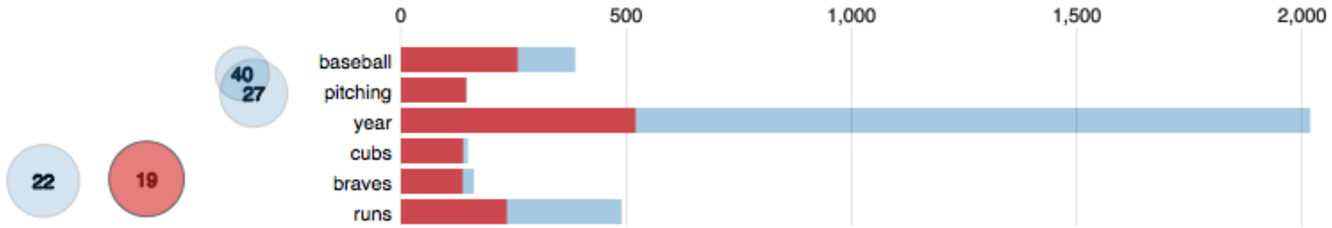
Topic #0	Topic #1	Topic #2	Topic #3	Topic #4
hand	looked	night	mr	elinor
good	found	room	miss	mother
madame	side	door	mrs	sister
life	speak	long	emma	marianne
heart	girl	house	jane	time
thought	gave	rochester	good	mrs
de	word	round	elizabeth	felt
day	made	hour	thing	letter
monsieur	sense	heard	dear	make
eye	eyes	back	great	john

Word clouds

Interpreting topic models via visualization

Interactive visualization

Visualizing topic models - pyLDAvis

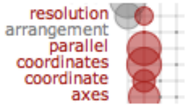


Interpreting topic models

More advanced

http://vis.stanford.edu/topic-diagnostics/model/silverStandards/

CITATION



Termite: Visualization Techniques for Assessing Textual Topic Models
Jason Chuang, Christopher D. Manning, Jeffrey Heer
Advanced Visual Interfaces, 2012
PDF (2.3 MB) | Website | Software

Distinctiveness & Saliency

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

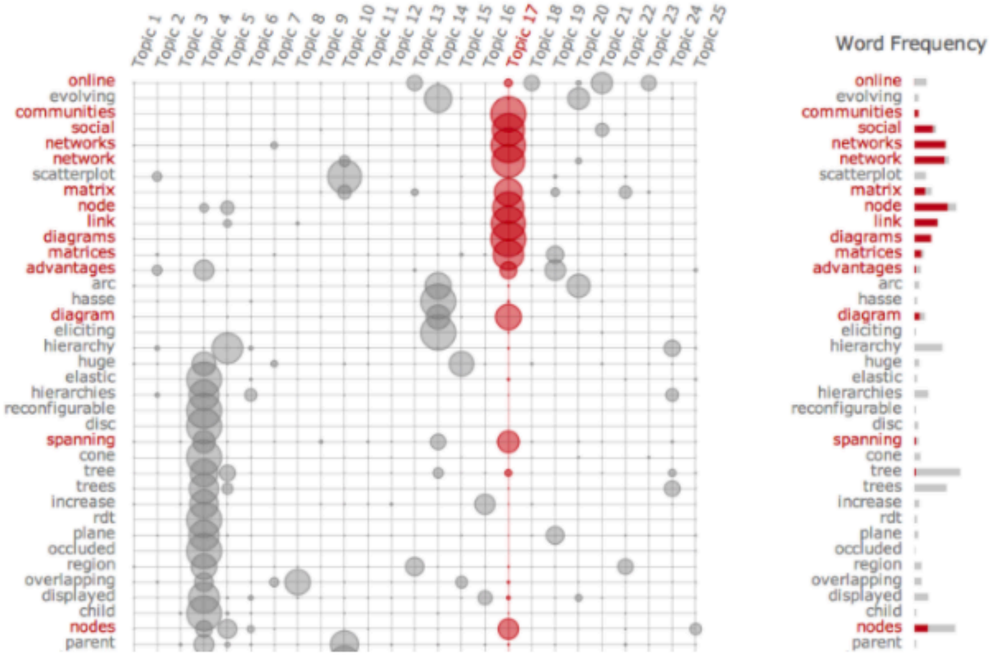
$$saliency(w) = P(w) \times distinctiveness(w)$$

We moved to Seattle! We packed our bags and headed north to become the University of Washington **Interactive Data Lab**. Come visit us...

STANFORD VIS GROUP HOME PAPERS PEOPLE VIDEO

Termite: Visualization Techniques for Assessing Textual Topic Models

Jason Chuang, Christopher D. Manning, Jeffrey Heer



The Termite system. A tabular view (left) displays term-topic distributions for an LDA topic model. A bar chart (right) shows the marginal probability of each term.

ABSTRACT

Topic models aid analysis of text corpora by identifying latent topics based on co-occurring words. Real-world deployments of topic models, however, often require intensive expert verification and model refinement. In this paper we present Termite, a visual analysis tool for assessing topic model quality. Termite uses a tabular layout to promote comparison of terms both within and across latent topics. We contribute a novel saliency measure for selecting relevant terms and a seriation algorithm that both reveals clustering structure and promotes the legibility of related terms. In a series of examples, we demonstrate how Termite allows analysts to identify coherent and significant themes.

Lab

Next Class - March 9th (Project Day)

1 - on -1 project group meetup.

Group #	Class Time (am PST)
1	8:30 - 8:40
2	8:40 - 8:50
3	8:50 - 9:00
4	9:00 - 9:10
5	9:10 - 9:20
6	9:20 - 9:30
7	9:30 - 9:40
8	9:40 - 9:50
9	9:50 - 10:00
10	10:00 - 10:10
11	10:10 - 10:20

Phase 3 submission - March 11th

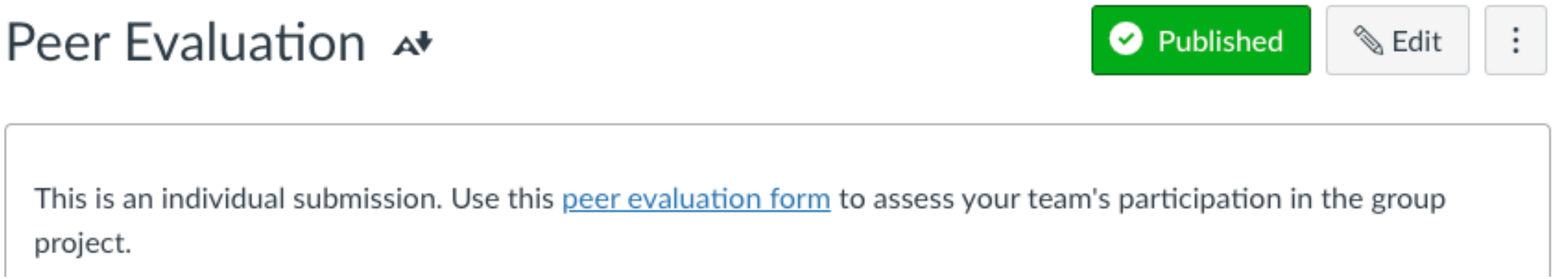
Videos due on March 11th before class (by **7:30am** PT) no late days.

Attendance mandatory on 11-March. All present in class for final presentation and Q&A

We will start at 8:30am sharp!!

Notebooks due by 11:59pm. No late days

Peer evaluations due by 5pm, March 12th. No late days



BREAK

Back at 9:45am

Readings

Journal
Information, Communication & Society ›
Volume 15, 2012 - Issue 5: A decade in Internet time: the dynamics of the Internet and society

CRITICAL QUESTIONS FOR BIG DATA

Provocations for a cultural, technological, and scholarly phenomenon

danah boyd & Kate Crawford

Big Questions for Social Media Big Data:
Representativeness, Validity and Other Methodological Pitfalls

Zeynep Tufekci
University of North Carolina, Chapel Hill
zeynep@unc.edu

Working with Big Data is still subjective

Boyd & Crawford paper

"This is exactly what I keep thinking during our project. It's only when I conducted the whole data science process in a project that I realized how subjective the analyses may be. **As a data analyst, I have to make a lot of choices, from data collection to cleaning, and these choices may impact the results with or without my awareness**."

Huan Wang

"there remains a mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts. In this way, Big Data risks re-inscribing established divisions in the long running debates about scientific method and the legitimacy of social science and humanistic inquiry. "

Boyd & Crowford paper

Bigger data are not always better data

> "We don't have better algorithms, we just have more data. More data beats clever algorithm, but better data beats more data." - Peter Norvig - Director of Research, Google

"Big vs whole data: In the project we are doing for the class, it is important to realise that the data that has been collected is not holistic….it is very likely that our dataset does not include several terms that are related to both topics and is therefore not whole in spite of being big."

- Anusha

"**Qualitative pull-outs**. Researchers can include a "qualitative pull-out" from their sample to examine variations in behavior. For example, what percent of retweets are "hate-retweets"? A small random subsample can provide a check. "

- Tufecki paper

Methodological considerations with social media data

Hashtag Analyses - The inclusion of hashtags in tweets is a Twitter convention for marking a tweet as part of a particular conversation or topic, and many social media studies rely on them for sample extraction.

Regarding the bias of hashtags, for a recent small project, it was almost impossible to find negative tweets from haters in the #blacklivesmatter movement.

Ayushi Gaur

Methodological considerations with social media data

Non-representativeness at the level of ***mechanisms*** as well as ***samples***.

"…take he example of the group project that my team is currently working on for this class - using Reddit data to analyze whether public discourse on domestic violence has changed since the onset of the COVID-19 pandemic. In this case, we know that Reddit is not representative of the global (or for that matter, even domestic American) population experiencing some form of domestic violence, so subreddits are essentially self-selecting sampling frames."

- Divya

Methodological considerations with social media data

Human self-awareness needs to be taken into account; humans will alter behavior because they know they are being observed, and this change in behavior may correlate with big data metrics.

- Tufecki paper

“ doubt whether users on the trending social media would change their own behaviors due to self-awareness. The reason I am being skeptical is because of my own experience on the major social media platforms, such as Instagram, Twitter, and TikTok. In general, I do not totally agree with point ..”

- Yuanfeng

Ethical concerns/considerations

“I feel a little bit guilty to analyze the data we accessed legally from Twitter and Reddit. The data do not contain users' sensitive personal data, but they were created by real living people. The words they posted on social media, whether or not they really mean it, we are analyzing their behavior through their words. I feel creepy when I imagine if my data had also been included in our analysis. That brings me to the ethics that although we legally have users’ consent to access their data and perform analysis on their data. Is it ethical for us to interpret the results based on the methodology we defined?”

- Esther Yang

Limitations, Ethical considerations for your projects

Work in groups to narrow this down

Project Phase 3 - notebook

report. Your notebook should also end with three additional sections (this can be written in markdown cells):

- Limitations: What are some limitations of your project as of now? You can rely on the readings assigned on the Ethics topic in class to talk about representativeness, validity, and other methodological pitfalls of your project.
- Ethical considerations: Do you foresee your project having any ethical concerns? As you did the readings on the Ethics topic, what are some ethical considerations that came to mind?
- Future work: How can you extend the current project (if at all). Another way to read this question is: If you had more time what would you have done?