Week 10 Agenda

- Announcements / house-keeping
- Intro to text-as-data
- Research example
- Live coding example
- Thursday preview

What is “text-as-data”?

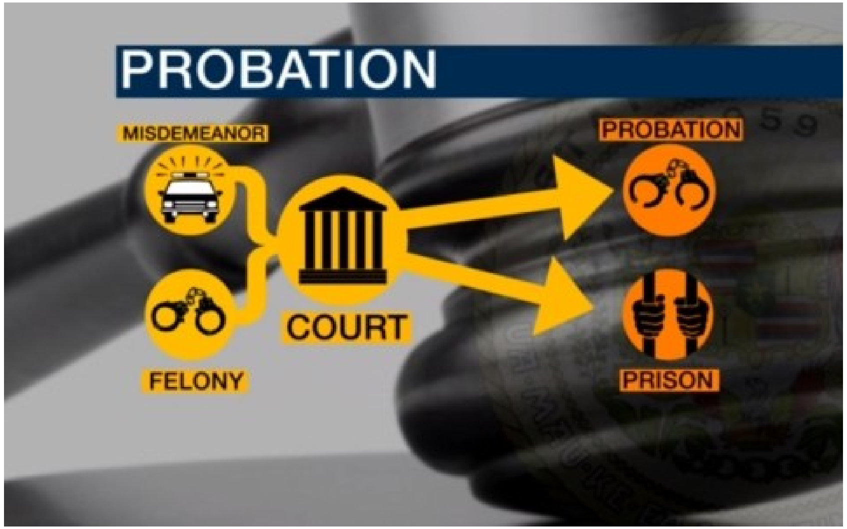
- “Information” can be encoded in lots of ways!
- “Text-as-data” is any attempt to **systematically** decode information from textual communication
- Often quantitative...
- ...but could be qualitative (“reading”).

Differences from “typical” data

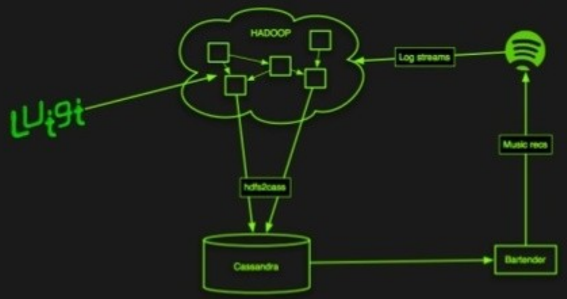
- Highly dimensional: e.g., corpus with *w* words drawn from *p* vocabulary = p^w dimensions!
- “Unstructured”: structured by linguistic rules, must be processed for analysis

REVOLUTION TOOLS
~~AK-47~~
~~MACHETE~~
TWITTER ✓
FACEBOOK ✓

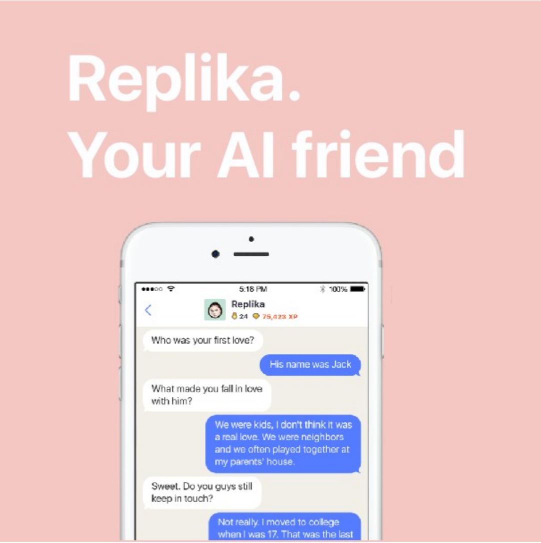
FAKE NEWS



The Discover page



Okay, but how do we come up with recommendations?
Collaborative filtering!



Who was your first love?

His name was Jack

What made you fall in love with him?

We were kids, I don't think it was a real love. We were neighbors and we often played together at my parents' house.

Sweet. Do you guys still keep in touch?

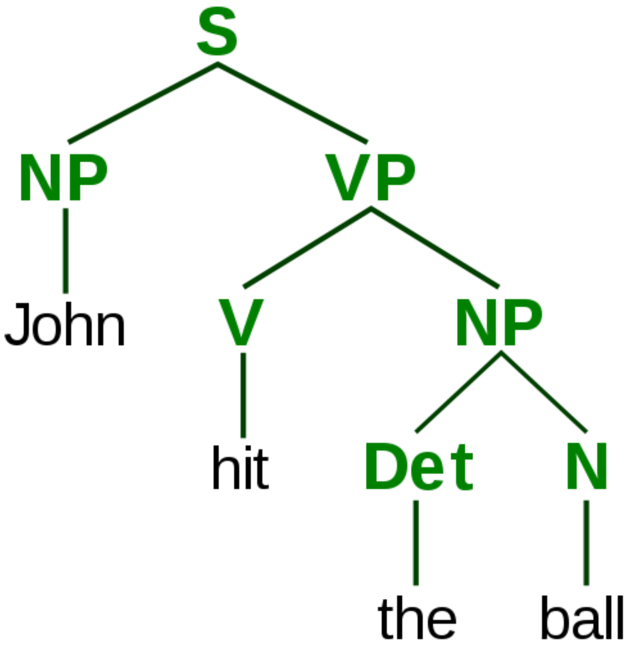
Not really. I moved to college when I was 17. That was the last

Brief history of text-as-data

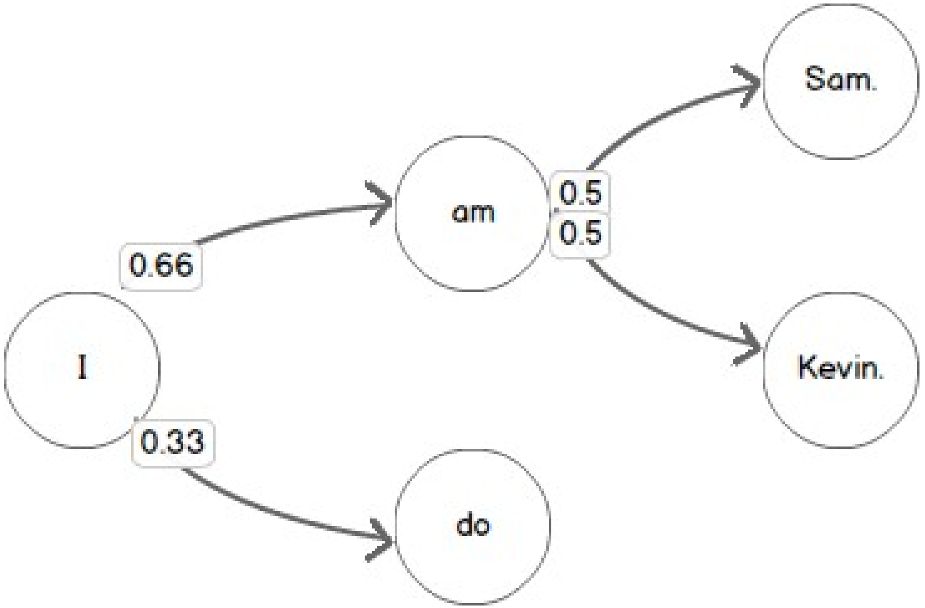
- **Harold Lasswell**: Political scientist(!) who “invented” quantitative content analysis
- 1934: the first (published) use of a key-word count

Brief history of text-as-data

- Chomskyan formalism



- **N**atural **L**anguage **P**rocessing (NLP)



Major tasks of NLP

- **Processing / feature exraction**
- Stemming
- Parts-of-speech tagging
- Sentence boundary disambiguation

- **Analysis**
- Word counts / dictionaries
- Classification
- Topic modeling

Processing

- “Processing” text produces **features:** individual measurable attributes of the text.
- These can include properties of the text itself, or metadata about the document or corpus.

Feature examples

- **Tf-idf**: term frequency, inverse document frequency. A measure of word “importance.”
- **N-grams:** a continuous sequence of *n* words. A measure of phrase frequency.
- **Adjacency matrix:** a matrix storing relative distances between words. A measure of co-occurance.

Analysis

- “Analysis” uses features and statistical tools to make inferences about texts or the corpus
- **Latent** vs. **manifest** characteristics:
    - **Latent:** inference about creator / source of text
    - **Manifest:** form & nature of communication
- Analytical tasks need not be complicated!

Analytical examples

- **Classification:** using features to identify which category a text belongs to (e.g., authorship, ideology)
- **Sentiment analysis:** using features to identify and extract affective information about a text (unreliable, basically astrology for data scientists)
- **Topic modeling:** text-mining for underlying semantic structures