Lecture 7 – Text Analysis
Part 2: Text Analysis

Prof. Walid Magdy, Informatics School

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

Scenario

- You are given access to a new dataset
    - 2 corpora, each contains thousands of text files
    - You want to understand and quantify:
        - What is the *content* of these documents? What are they *about*?
        - How does the content of these corpora *differ*?

- How can you analyse?

Lecture Objectives

- Analyze text corpora
  - Content analysis background
  - Word-level differences
  - Dictionaries and Lexicons
  - Topic modeling

Content Analysis

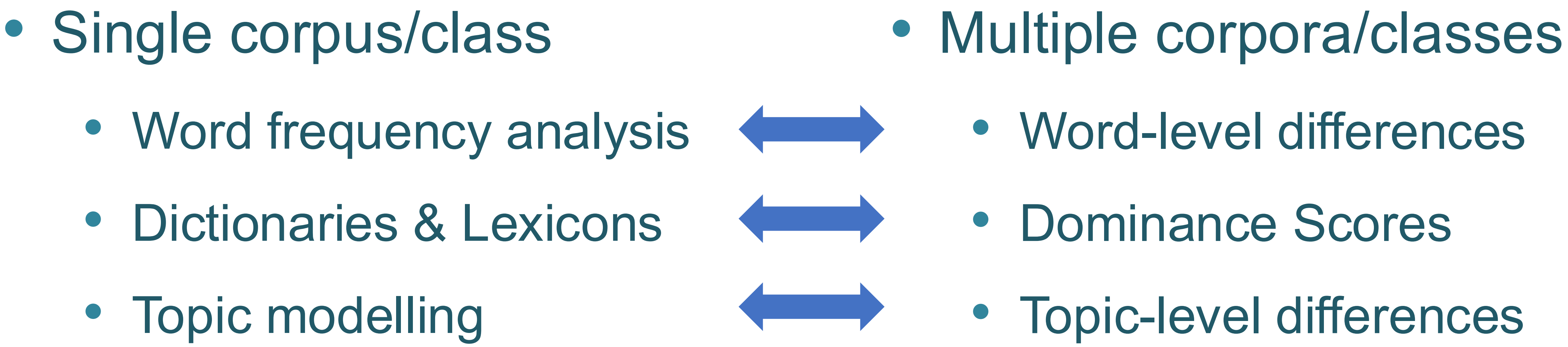
- Goal: given some documents determine
  - What are the types of content present? (themes/topics)
  - Which documents contain which topics?

- Traditionally a manual process
  1. Read a subset of documents, define themes/topics
  2. Determine consistent ***thematic coding*** methodology
  3. Read all documents and label them according to codes
  4. Check agreement between human coders
  5. Settle disagreements via a third-party
  6. Analyze resulting annotations

Content Analysis

- Can this process be automated?
    - Yes, to an extent
- *Should* this process be automated?
    - Humans are better than machines at this task (for now?)
    - Computers are *much*, *much* faster
        - Avg. human reading speed: 250 wpm
        - Assume 1K words/document, 50K documents…
            - Average person needs > 4 months to read
            - This is a ***relatively small*** corpus for modern NLP
        - Modern computers can process millions of words/second

Automated Content Analysis



- Single corpus/class
  - Word frequency analysis
  - Dictionaries & Lexicons
  - Topic modelling

- Multiple corpora/classes
  - Word-level differences
  - Dominance Scores
  - Topic-level differences

Word Level Analysis

Word frequency analysis

- Very simple starting point

1. Preprocess as usual (lowercasing? stemming?)
2. Count words
3. Normalize by document length
4. Average across all documents

Word-level Differences

- Word frequency for comparing two corpora?
  - Any issues?

- Which words best characterize a corpus?
  - Need a reference corpus

- Some methods to do this:
  - Mutual information
  - Chi squared

Mutual Information (MI)

- I(X;Y)
  - How much can I learn about X by observing Y?
  - Is the same as *information gain*
  - Is **not** the same as *pointwise mutual information*
- We want to learn about important words in our corpus
- What should X and Y be?
  - X = U = document contains term t (Boolean)
  - Y = C = class (group) is the target class (Boolean)

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t) P(C = e_c)}$$

$$\begin{aligned} I(U;C) = {} & \frac{N_{11}}{N} \log_2 \frac{N N_{11}}{N_{1.} N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{N N_{01}}{N_{0.} N_{.1}} \\ & + \frac{N_{10}}{N} \log_2 \frac{N N_{10}}{N_{1.} N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{N N_{00}}{N_{0.} N_{.0}} \end{aligned}$$

Chi-squared (x^2)

- Hypothesis testing approach
- H_0: Term appearance is independent from a document's class
  - i.e., P(U=1,C=1) = P(U=1)P(C=1)
- Compute:

$$X^2(\mathbb{D},t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

- Or to directly plug in values like before:

$$X^2(\mathbb{D},t,c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11} N_{00} - N_{10} N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

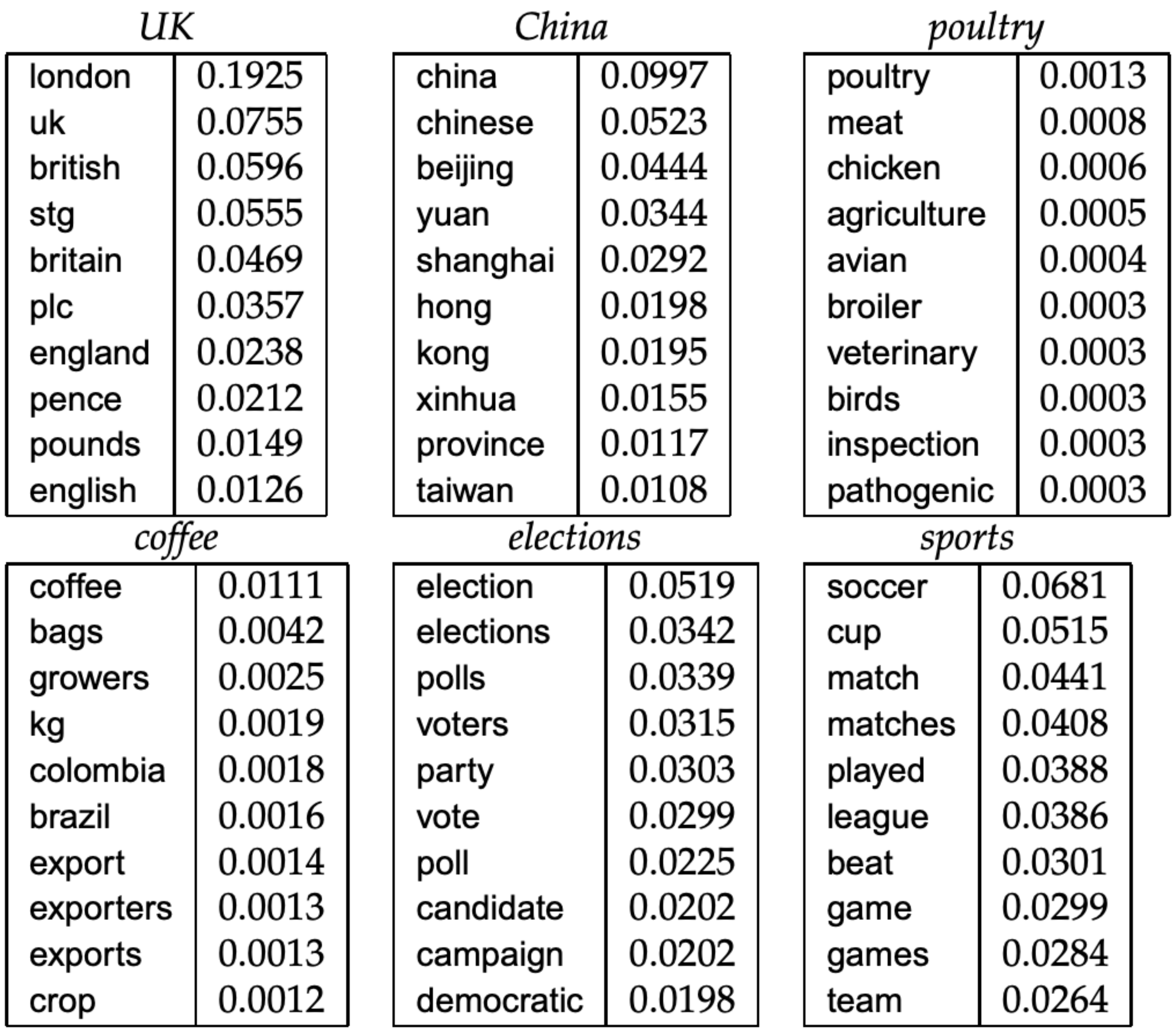
MI and x^2 , in English

What terms (features) are special (distinctive) in this group compared to the other groups?

Source: Manning, Raghavan, and Schütze, 2008

Example 1: **MI for News Data**

Example:
Manning, Raghavan, and Schütze, 2008



UK

london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China

china	0.0997
chinese	0.0523
beijing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

elections

election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

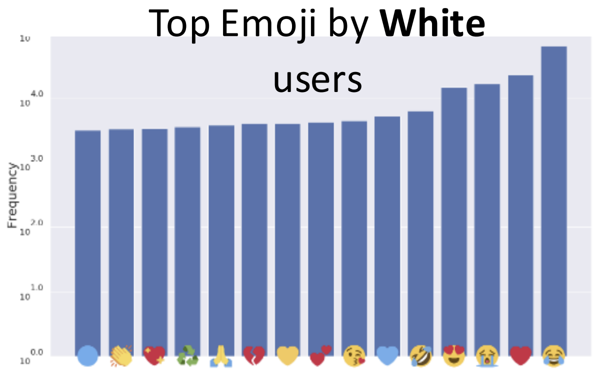
sports

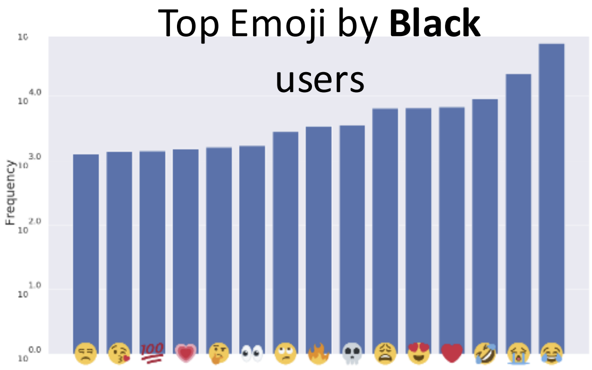
soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

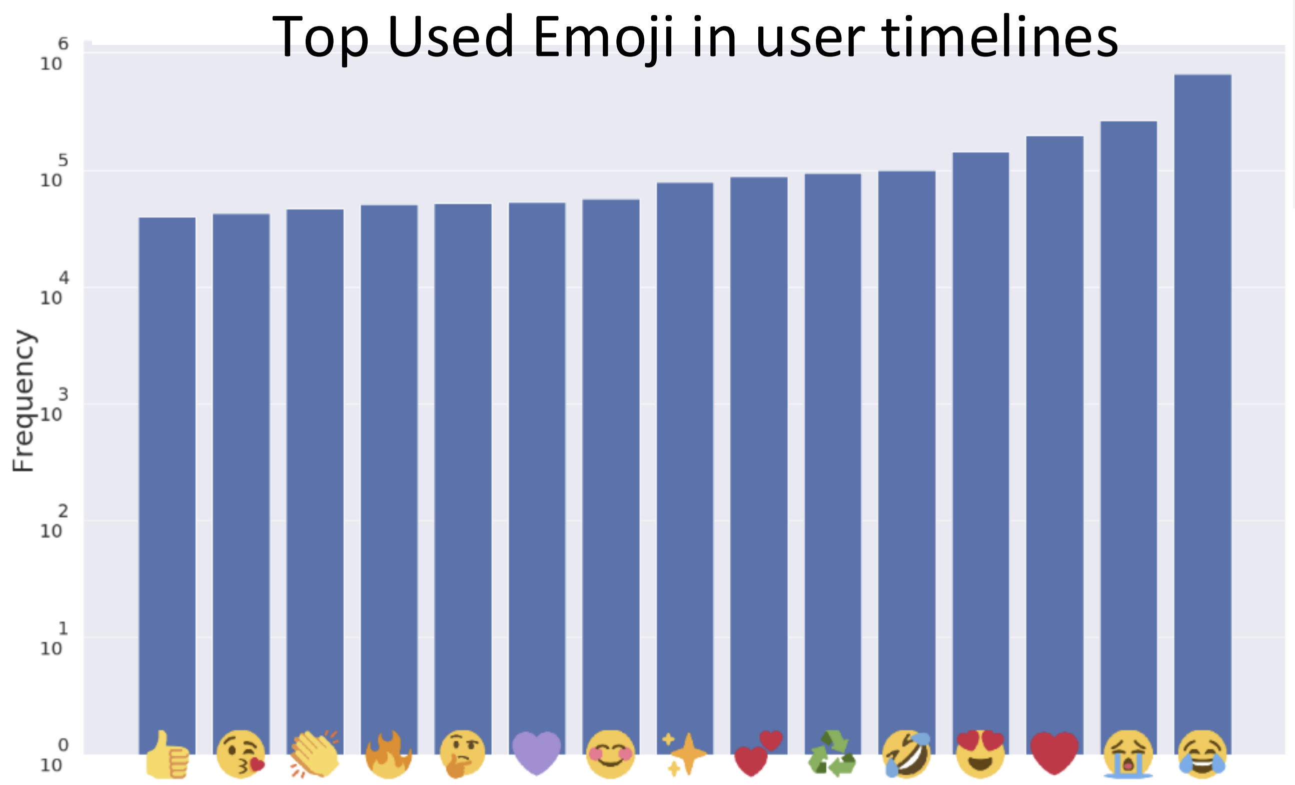
Example 2: **Emoji usage** ☺

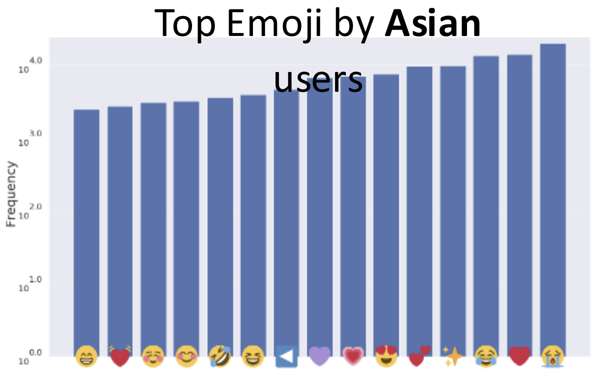
- **20+%** of social media posts contain emoji
- Does general emoji usage differ by demographic?
  - Collected timelines of **20,000** Twitter users (WW, NYC, London, Johannesburg)
  - Annotate ethnicity and gender of user based on profile pic.

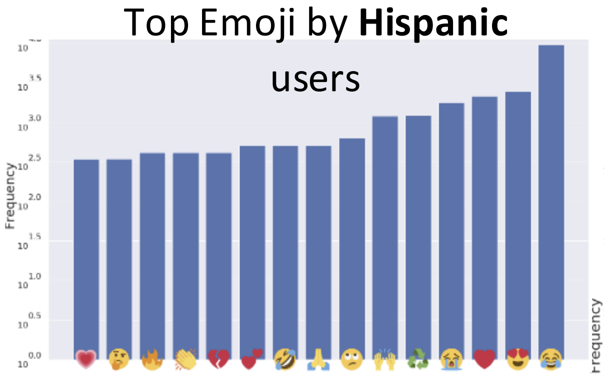
Most Frequent Emoji

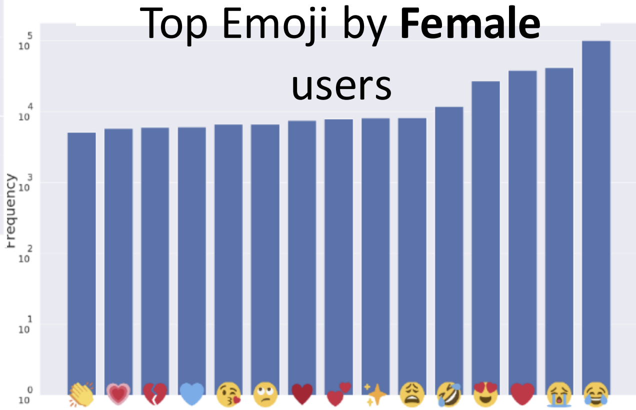

Top Emoji by **White** users

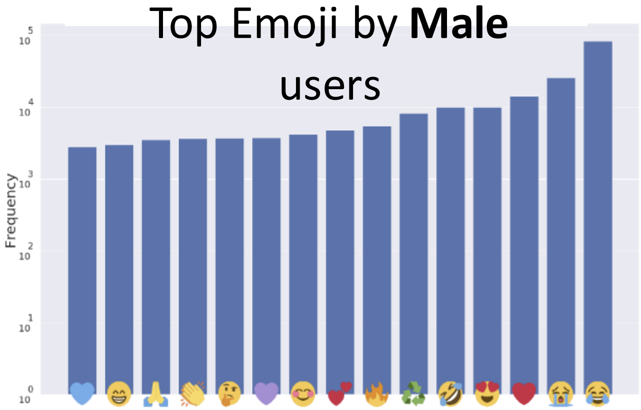

Top Emoji by **Black** users


Top Used Emoji in user timelines


Top Emoji by **Asian** users


Top Emoji by **Hispanic** users

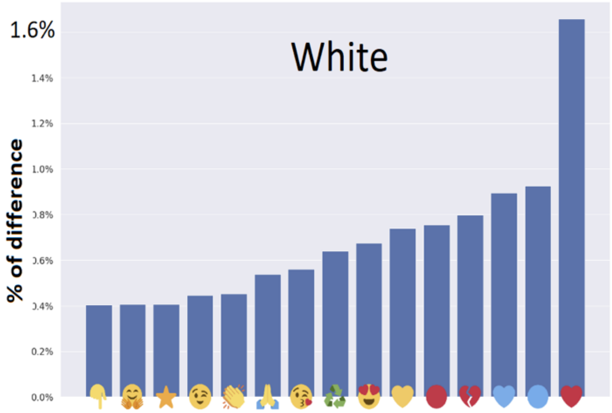

Top Emoji by **Female** users

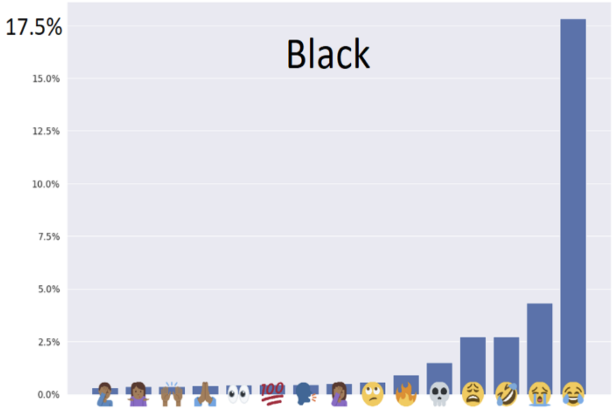

Top Emoji by **Male** users

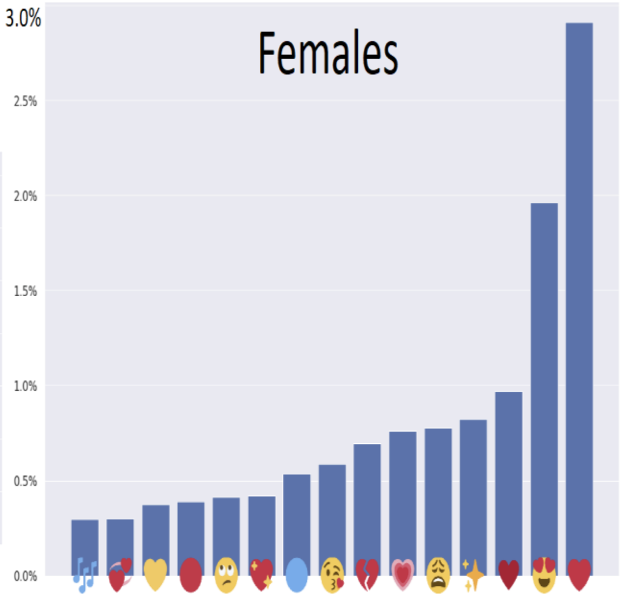
THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

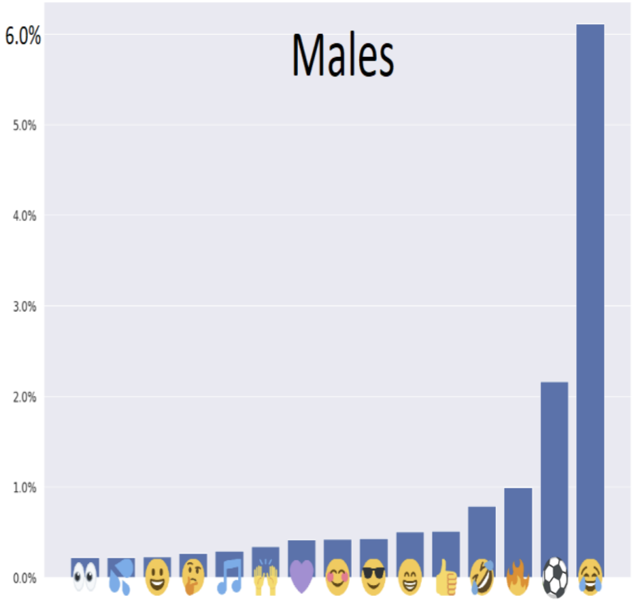
THE UNIVERSITY *of* EDINBURGH
informatics

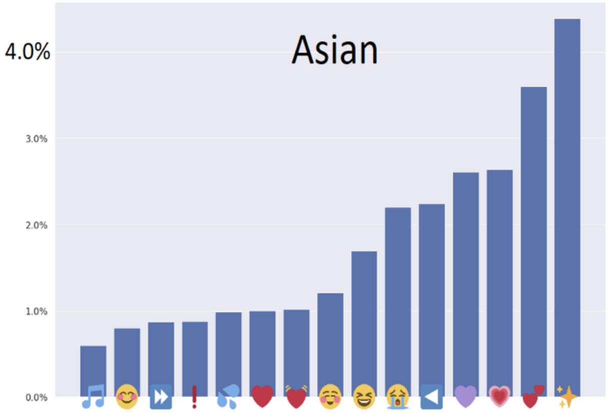
Applying Chi-squared

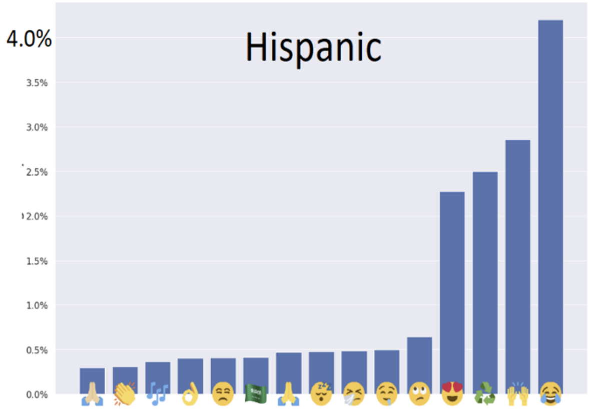












Dictionaries and Lexicons

Dictionaries and Lexicons

- What if we know what we are looking for?
- Dictionaries (lexicons) are prebuilt mappings
  - Category -> word list
  - E.g., a tiny sentiment lexicon:
    - Positive: good, great, happy, amazing, wonderful, best, incredible, healthy
    - Negative: terrible, horrible, bad, awful, nasty, gross, worst, poor, ill, sick

- Domain can be important
  - *“My son is **sick**”* ✓
  - *“The Movie was **Sick!**”* ✗

Dictionaries and Lexicons

- How to get a score per category?

$$\frac{num\_dictionary\_words\_in\_document}{num\_total\_words\_in\_document}$$

- That’s it!

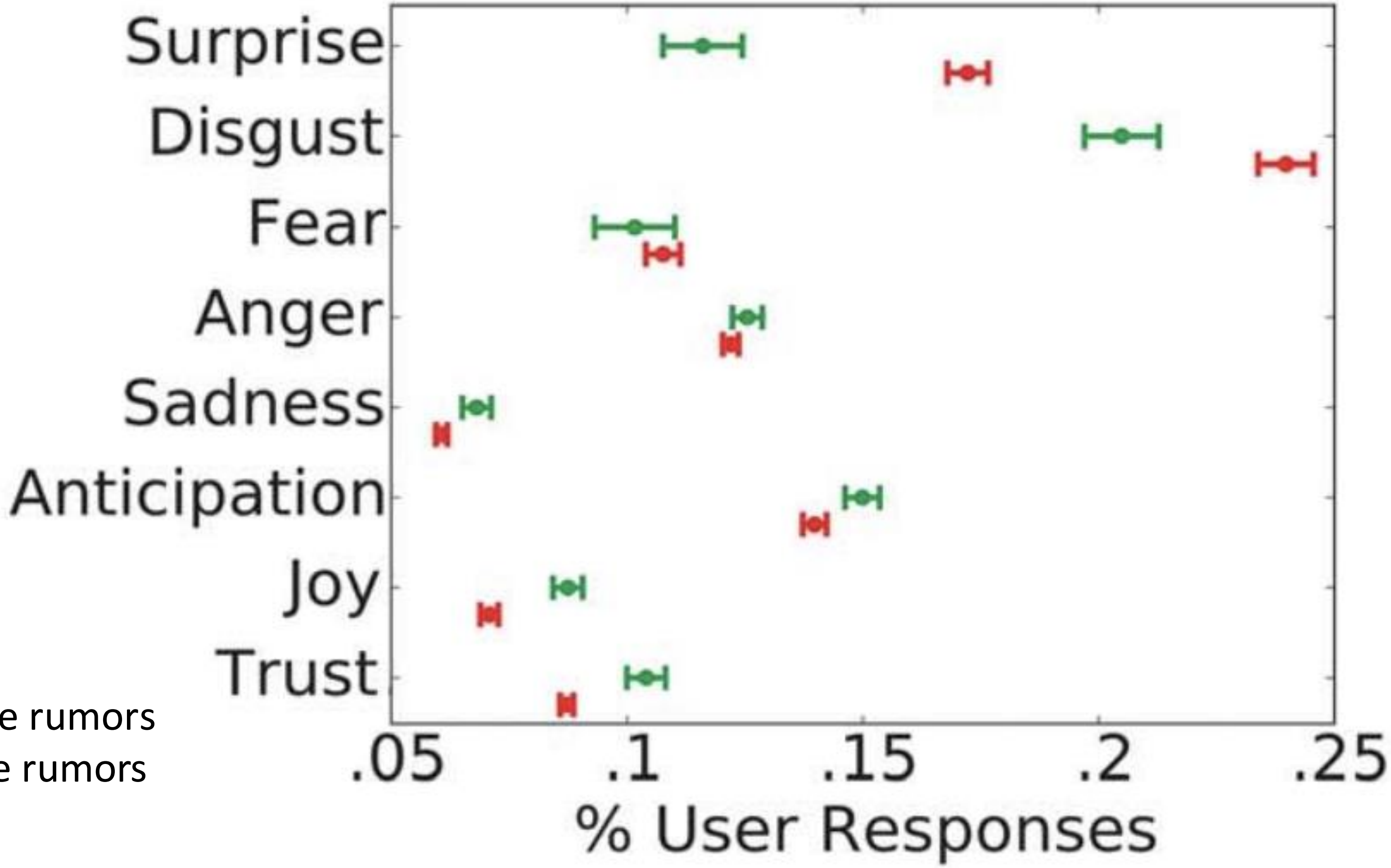
- Can also be used as machine learning features

- A more advanced approaches to quantifying categories (optional reading)
  - https://www.ncbi.nlm.nih.gov/pubmed/28364281

Some Dictionaries

• LIWC	(Pennebaker et al. 2022)
• General Inquirer	(Stone 1997)
• Roget’s Thesaurus Categories	
• VADER	(Hutto and Gilbert, 2014)
• Sentiwordnet	(Esuli and Sebastiani 2006)
• Wordnet Domains	(Magnini and Cavaglia, 2000)
• EmoLex	(Mohammad and Turney, 2010)
• Empath	(Fast et al., 2016)
• Personal Values Lexicon	(Wilson et al., 2018)
• …	

Example: Reactions to Rumor Tweets with EmoLex



Red = reactions to false rumors

Green = reactions to true rumors

Vosoughi, Roy, and Aral, 2018

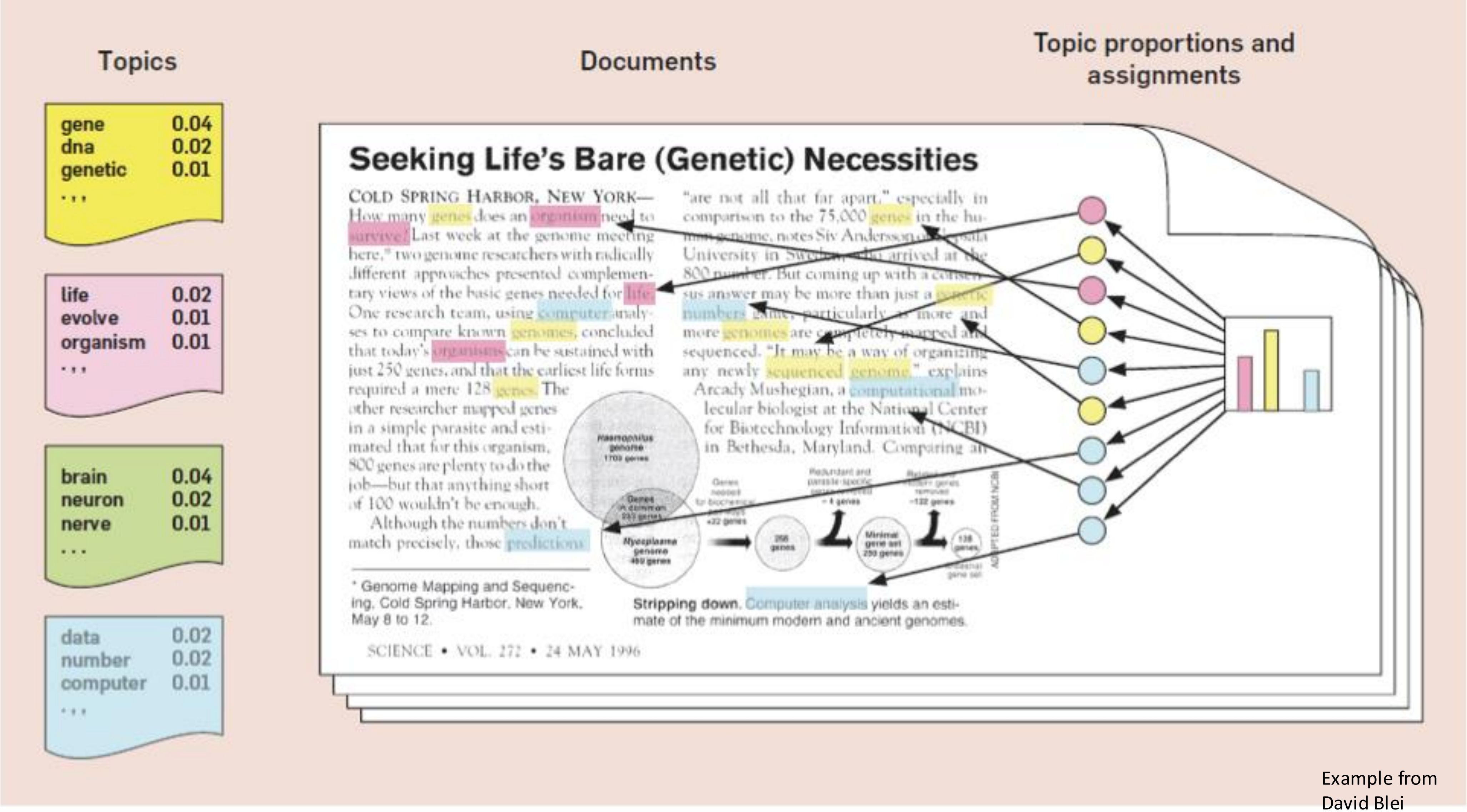
Topic Level Analysis

Topic Modelling

- Given a collection of documents
  → What are the list of topics discussed in these documents?

- Input: collection of documents
- Output: clusters of repetitive words to *N* topics discussed

- Most popular algorithms for topic modeling:
  - LDA (Latent Dirichlet allocation) → use surface form of words
  - BERTopic → use word embeddings carrying word meanings



Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

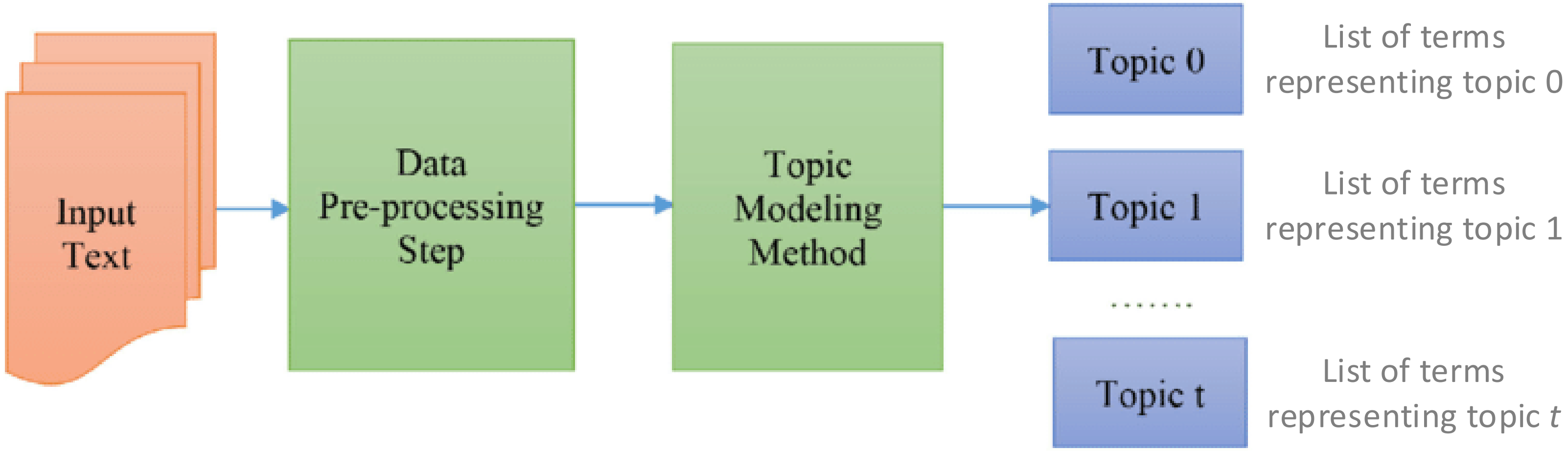
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Example from
David Blei

Topic Modelling



Input Text

Data Pre-processing Step

Topic Modeling Method

Topic 0

List of terms representing topic 0

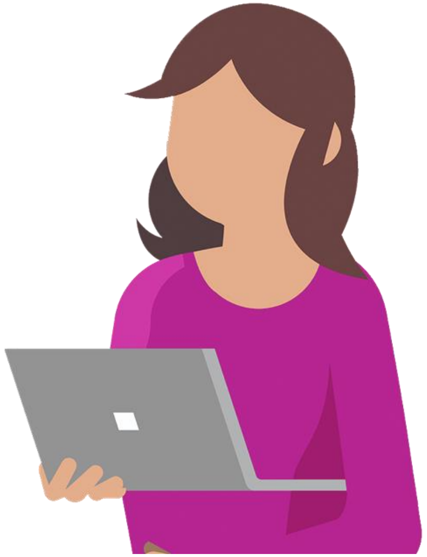
Topic 1

List of terms representing topic 1

....

Topic t

List of terms representing topic *t*



Manually name each topic

Readings

- LIWC-22: https://www.youtube.com/watch?v=IGBI8LnYGNs