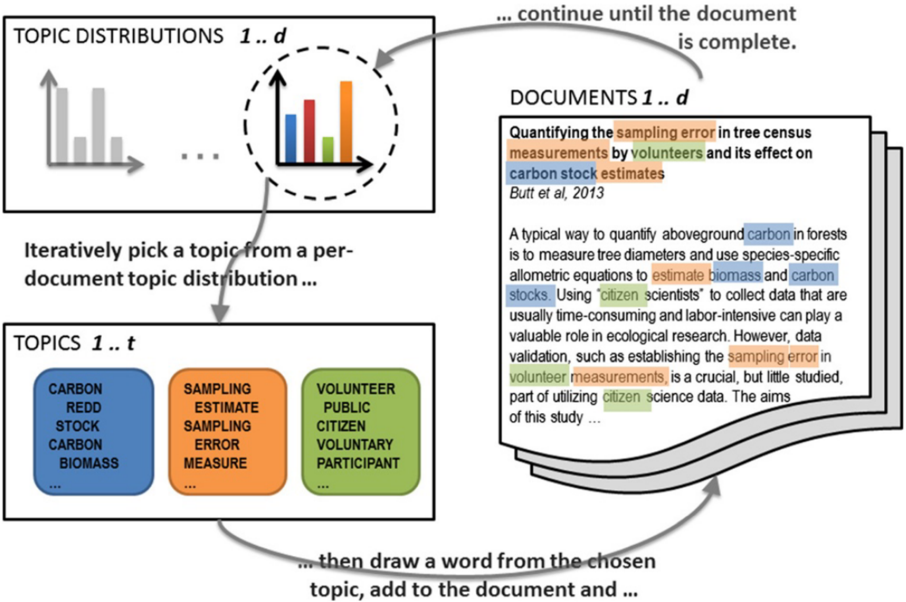
ASSUMED DOCUMENT GENERATION PROCESS



TOPIC DISTRIBUTIONS *1 .. d*

Iteratively pick a topic from a per-document topic distribution ...

TOPICS *1 .. t*

CARBON
REDD
STOCK
CARBON
BIOMASS
...

SAMPLING
ESTIMATE
SAMPLING
ERROR
MEASURE
...

VOLUNTEER
PUBLIC
CITIZEN
VOLUNTARY
PARTICIPANT
...

... then draw a word from the chosen topic, add to the document and ...

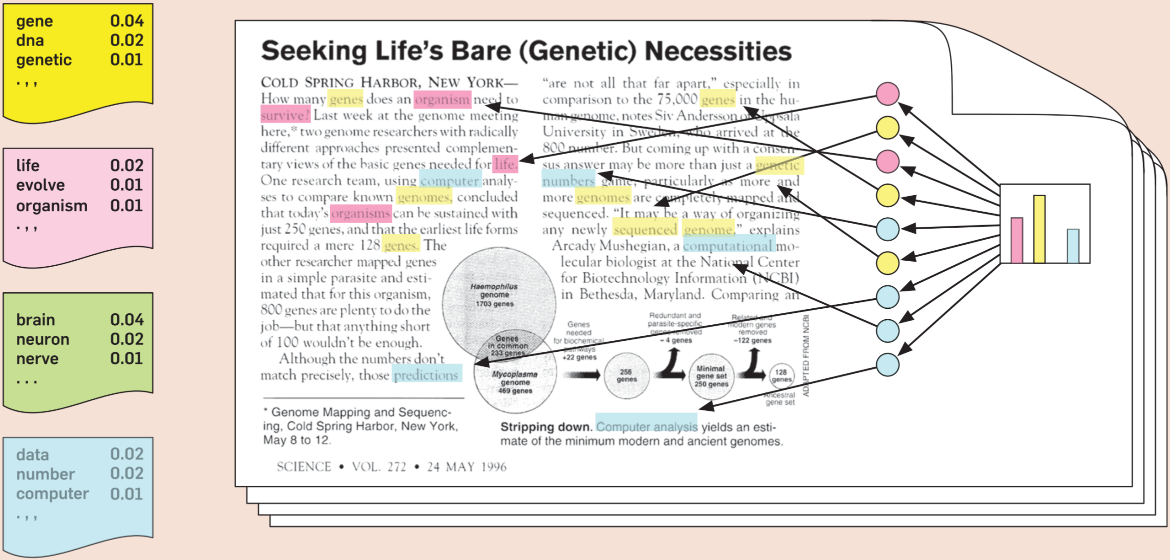
DOCUMENTS *1 .. d*

Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates

Butt et al, 2013

A typical way to quantify aboveground carbon in forests is to measure tree diameters and use species-specific allometric equations to estimate biomass and carbon stocks. Using "citizen scientists" to collect data that are usually time-consuming and labor-intensive can play a valuable role in ecological research. However, data validation, such as establishing the sampling error in volunteer measurements, is a crucial, but little studied, part of utilizing citizen science data. The aims of this study ...

... continue until the document is complete.



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

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 and 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

Topic proportions and assignments

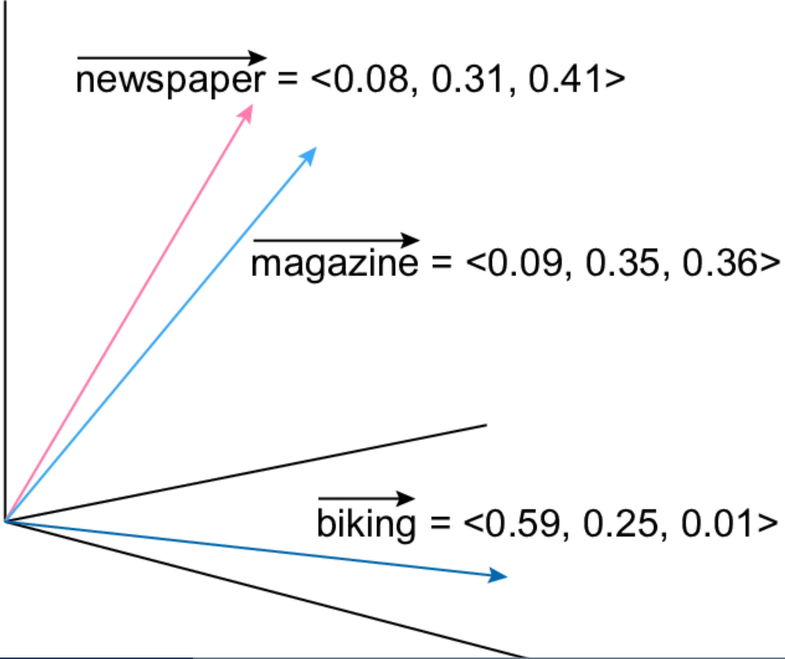
d documents and t topics (set in advance); model will then create $d + t$ probability distributions:

$$\Pr_d(\text{topic})$$

(How likely is each topic to appear in the given document?
Or, more informally, what's the mixture of topics in each document?)

$$\Pr_t(\text{word})$$

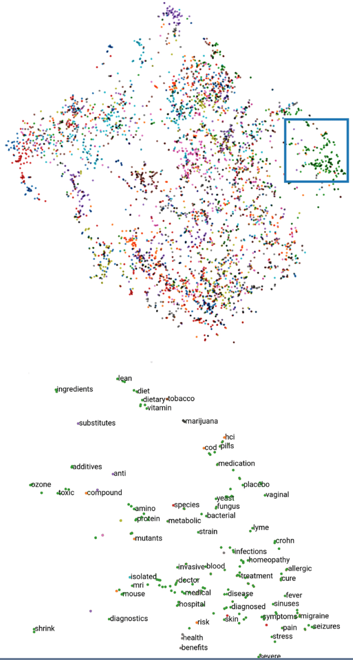
(How likely is each word to appear in each topic?
Or, more informally, what words "pick out" or "are important for" each topic?)



$\overrightarrow{\text{newspaper}} = <0.08, 0.31, 0.41>$

$\overrightarrow{\text{magazine}} = <0.09, 0.35, 0.36>$

$\overrightarrow{\text{biking}} = <0.59, 0.25, 0.01>$


lean
ingredients
diet
dietary
tobacco
vitamin
substitutes
marijuana
hci
cod
pills
medication
additives
anti
ozone
toxic
compound
placebo
vaginal
yeast
fungus
amino
species
bacterial
protein
metabolic
strain
lyme
mutants
crohn
infections
homeopathy
invasive
blood
allergic
isolated
treatment
cure
doctor
mri
mouse
medical
disease
fever
sinuses
hospital
diagnosed
symptoms
migraine
diagnostics
risk
skin
pain
seizures
shrink
health
stress
benefits
severe

Tried word embeddings using the Stanford GloVe pre-trained dataset, and got surprisingly useless results.

Hypothesis: It's confused by all the scientific/philosophical/etc. terminology. Could try using a model trained on scientific corpora like SciBERT, but I didn't have time!

Whichever tool you use:

1. Use the “distance” between documents (either their mix of topics, or each document’s average position in the word-vector space) to determine similarity.
2. Start with a randomly seeded conference schedule.
3. Randomly swap talks, using simulated annealing to get to a (hopefully) optimal schedule.

Things we can talk about:

- **DH Methods:** Interested in topic modeling or word embeddings? These things are useful all over the place.
- **Tech Details:** Interested in how the actual code that I wrote works? It's in Python.
- **Conference Scheduling:** What worked and what didn't? What did we learn?

Questions?

charles@charlespence.net
https://pencelab.be
@pencechp · @pencelab