A Practical Algorithm for Topic Modling with Provable Guarantees

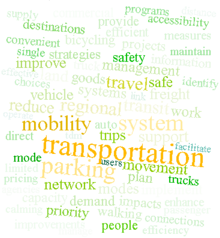
Sanjeev Arora
Rong Ge
Yoni Halpern
David Mimno
Ankur Moitra
David Sontag
Yihcen Wu
Michael Zhu

Presented by: Vanush Vaswani and Kristy Hughes

Information Overload

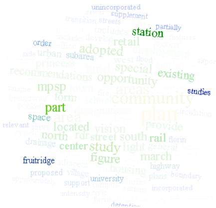
Effective Organisation

Topics

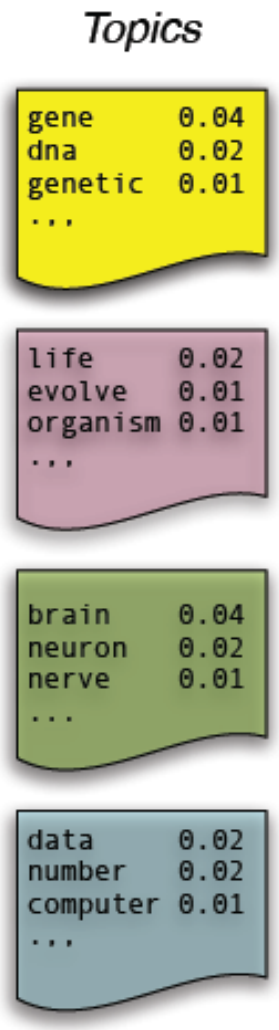

transportation
parking
system
mobility
travel
safe
regional
transit
movement
trucks
network
mode
trips
safety
management
people
priority


workforce
innovation
education
industry
economy
research
training
technology
systems
clusters
institutions
improving
metropolitan
entrepreneurs
skills


commercial
industrial
residential
areas
retail
mixed
activity
district
office
districts
environmental
concentration
surrounding


plan
community
area
station
part
study
rail
adopted
existing
located

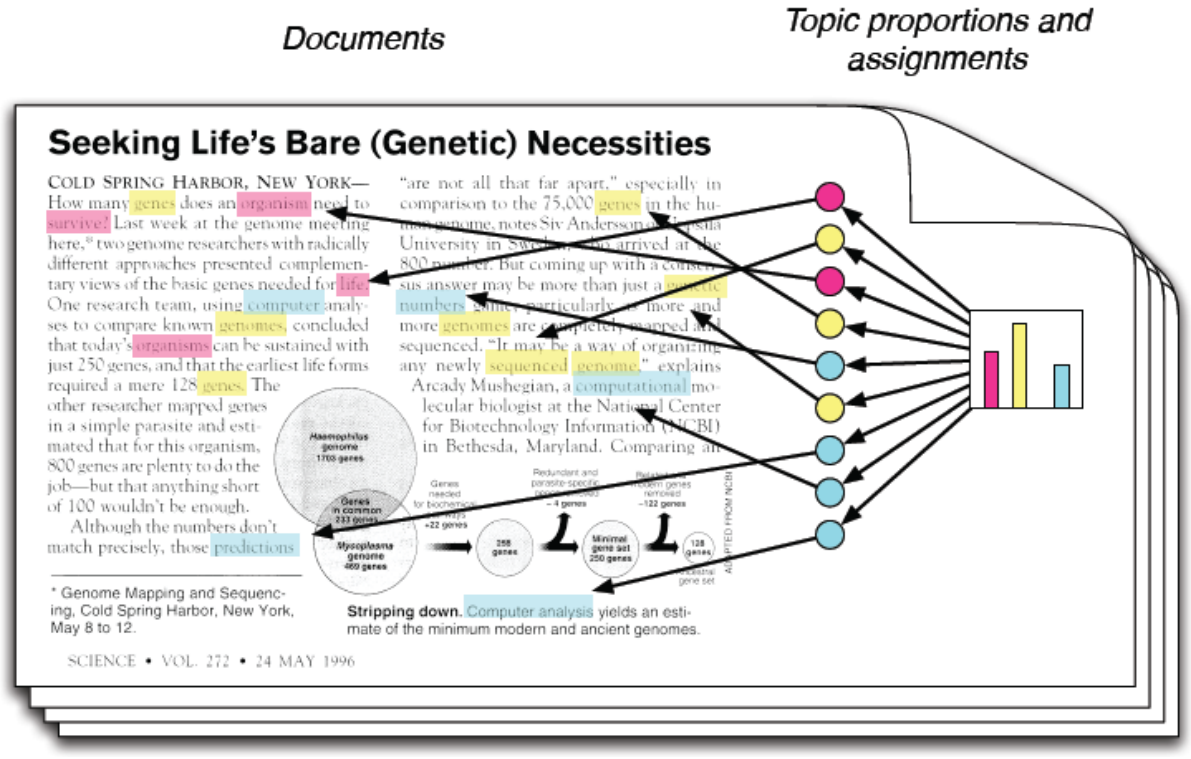
Model of Topics



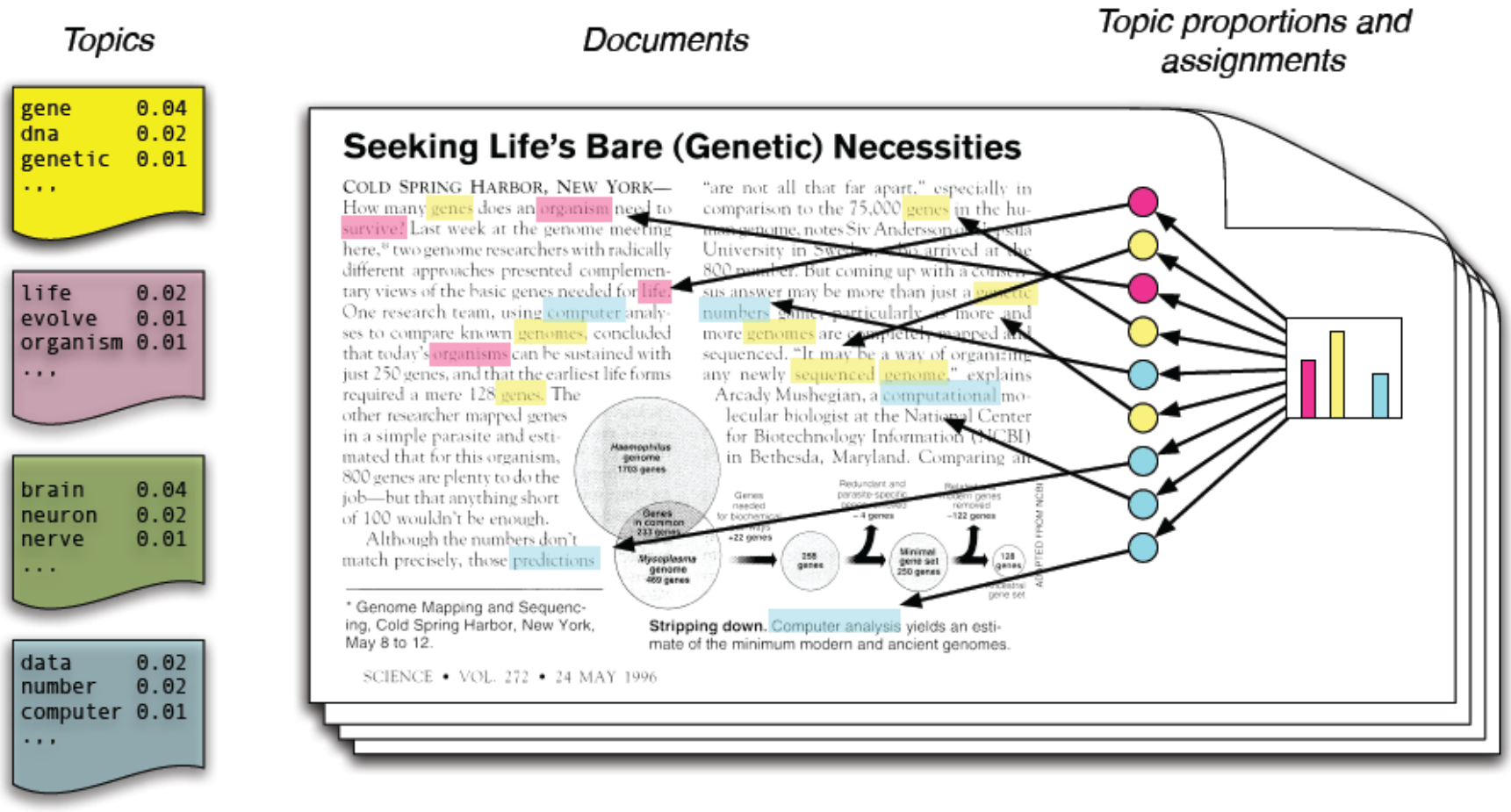
Topics are distributions over words

Model of Documents

Documents have distribution of topics

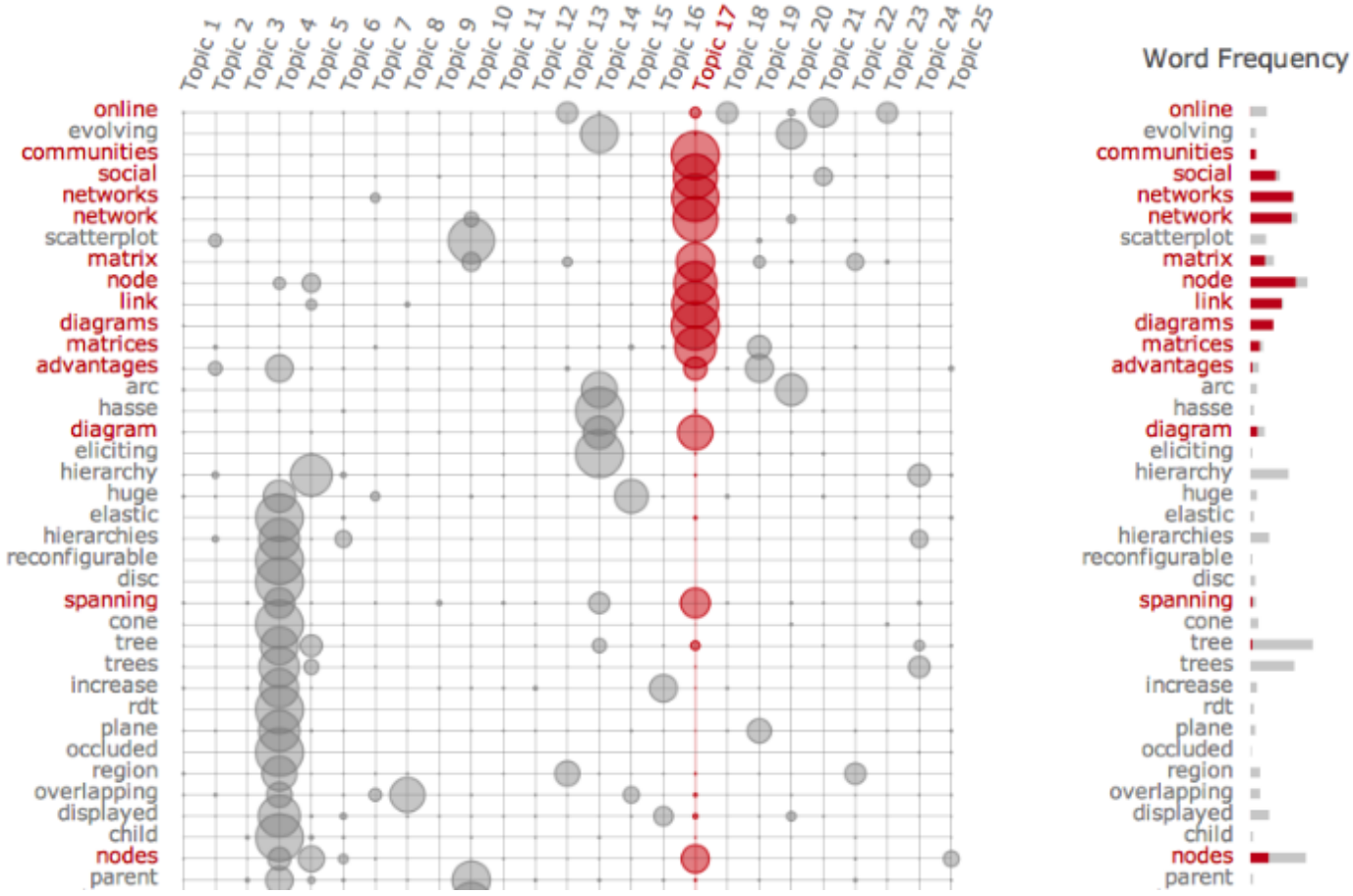


Topic Modelling



Word-topic Matrix

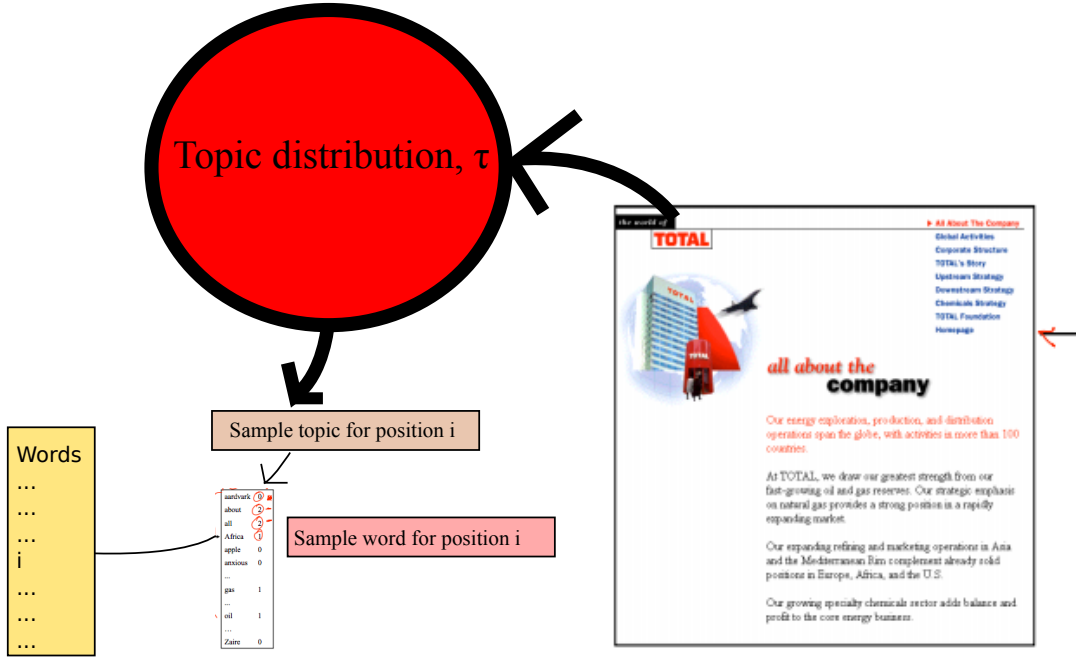
Extracted: Word-topic matrix



Aim: Find document-topic matrix

Steps

- Assume documents are generated by probabilistic model with unknown variables
- Infer hidden structure onto document
- Situate new document into model



TODO: Redo pic

Approximate Inference & Provable Guarantees

- Document-topic inference is NP-hard.
- Approximate techniques used
- Need provably polynomial-time algorithms

TODO: Draw pic

Algorithm

Steps

1. Second order moment matrix of word-word co-occurrences
2. Anchor word selection
3. Topic distribution recovery

Assumptions:

- Topics may be correlated
- Word-topic distributions are separable

Anchor Words

LDA

Contributions