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

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

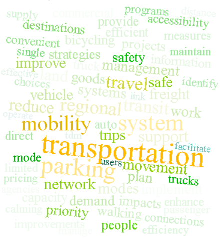
6 Experimental Results

7 Conclusion

Information Overload

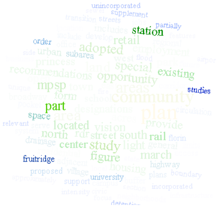
Effective Organisation

Topics


transportation
parking
mobility
system
movement
travel
safe
safety
network
trucks
mode
trips
regional
transit
management
people
priority
demand
programs
accessibility
destinations
convenient
bicycling
improve
strategies
goods
vehicle
freight
reduce
work
plan
direct
facilitate
impacts
walking
connections
efficiency
capacity
modes
support


workforce
innovation
education
economy
research
training
technology
industry
industries
clusters
systems
skills
institutions
improving
metropolitan
entrepreneurs
region
technologies
universities
important
skilled
work
data
future
educated
logistics
workers
employers
venture
talent
foundations


commercial
industrial
residential
retail
areas
mixed
activity
district
districts
existing
office
land
services
surrounding
shopping
designated
environmental
concentration
distribution
expansion
business
structures
housing
large
area
low
goods
location
oriented
accommodate
businesses
adjacent
proximity
serve
light
restaurants
located


community
plan
area
areas
station
adopted
retail
special
existing
opportunity
recommendations
mpsp
part
located
vision
street
south
rail
north
study
center
figure
march
housing
urban
subarea
space
studies
fruitridge
proposed
university
highway

1 Introduction

2 Topic Modelling

3 Algorithm

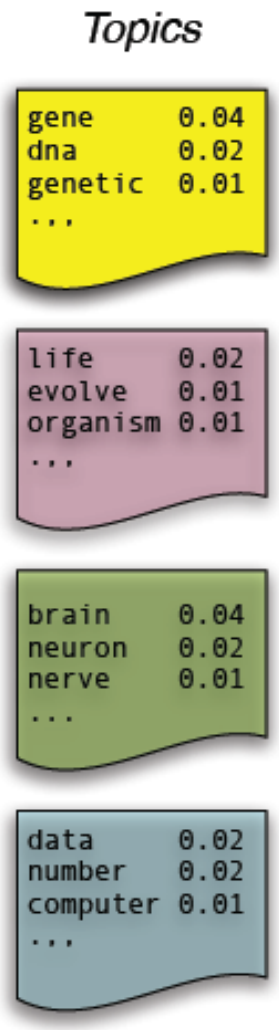
4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion

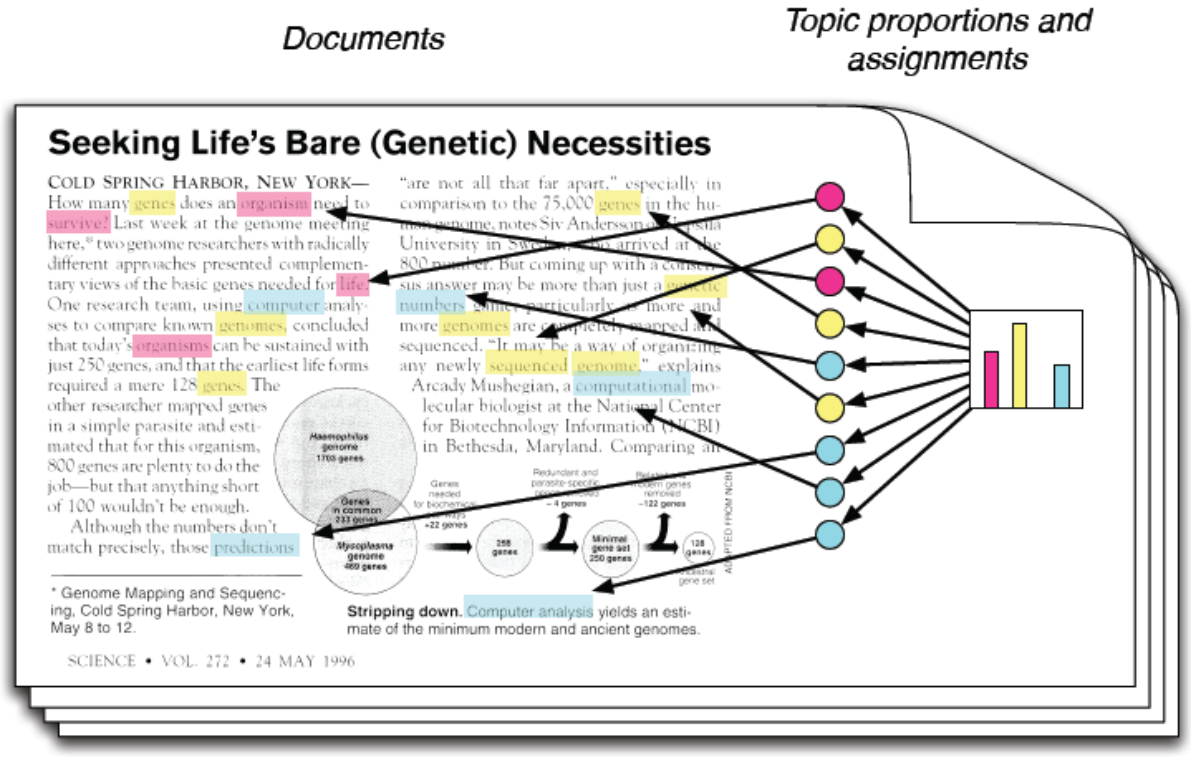
Model of Topics



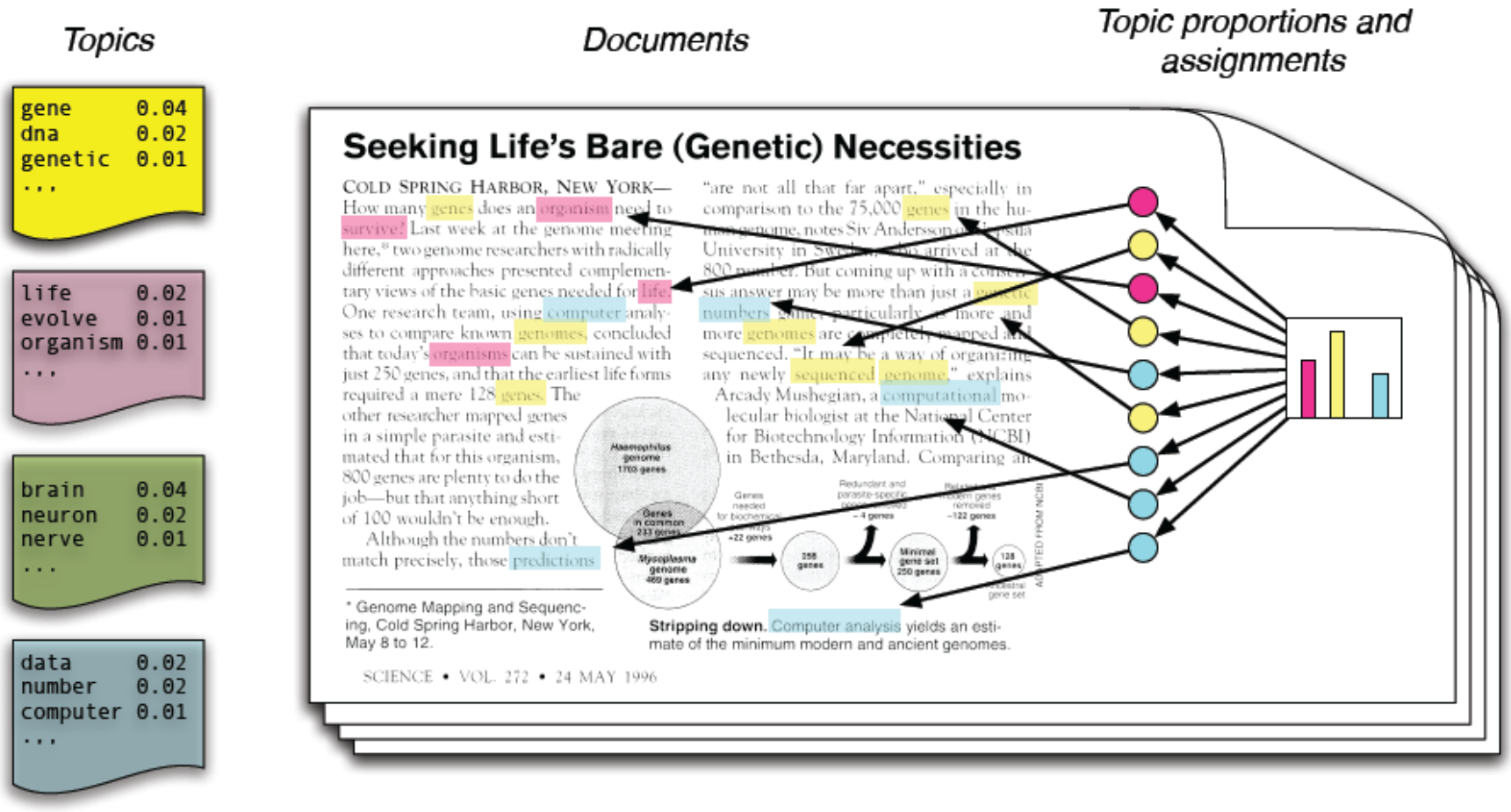
Topics are distributions over words

Model of Documents

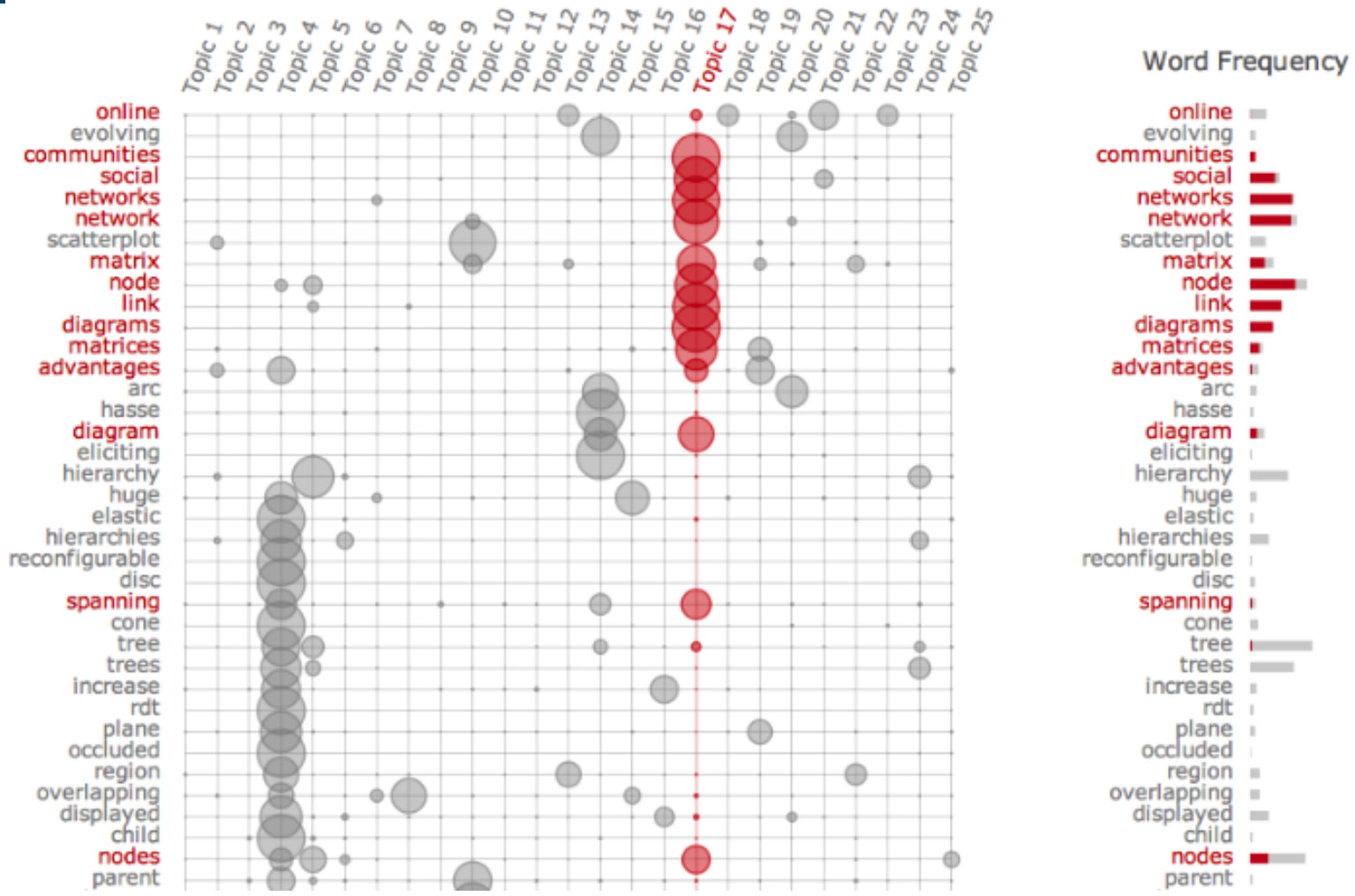
Documents have distribution of topics



Topic Modelling



Word-topic Matrix



Aim: Find document-topic model

Steps

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

Approximate Inference & Provable Guarantees

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

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion

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

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion

1 Introduction

2 Topic Modelling

3 Algorithm

4 Topic Recovery via Bayes' Rule

5 Efficiently Finding Anchor Words

6 Experimental Results

7 Conclusion