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

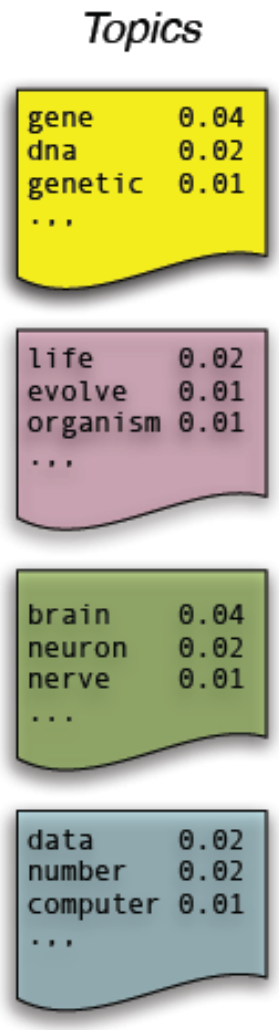
Information Overload

Effective Organisation

Topics


transportation
parking
mobility
system
travel
safe
safety
regional
transit
movement
network
trucks
mode
trips
priority
people
destinations
accessibility
programs
workforce
innovation
education
industry
training
technology
economy
research
institutions
systems
clusters
skills
universities
metropolitan
entrepreneurs
commercial
industrial
residential
retail
areas
mixed
activity
district
existing
office
environmental
concentration
designated
structures
large
surrounding
accommodate
plan
community
station
adopted
special
existing
opportunity
part
located
study
rail
figure
fruitridge

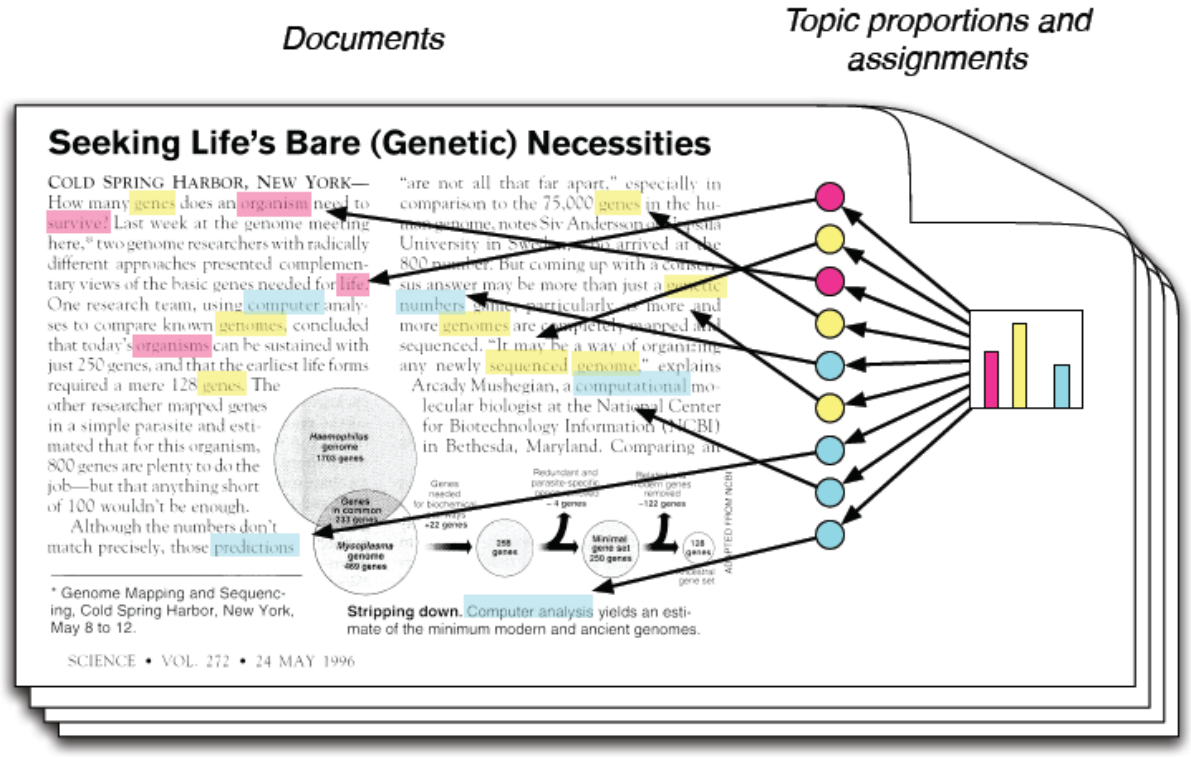
Topics



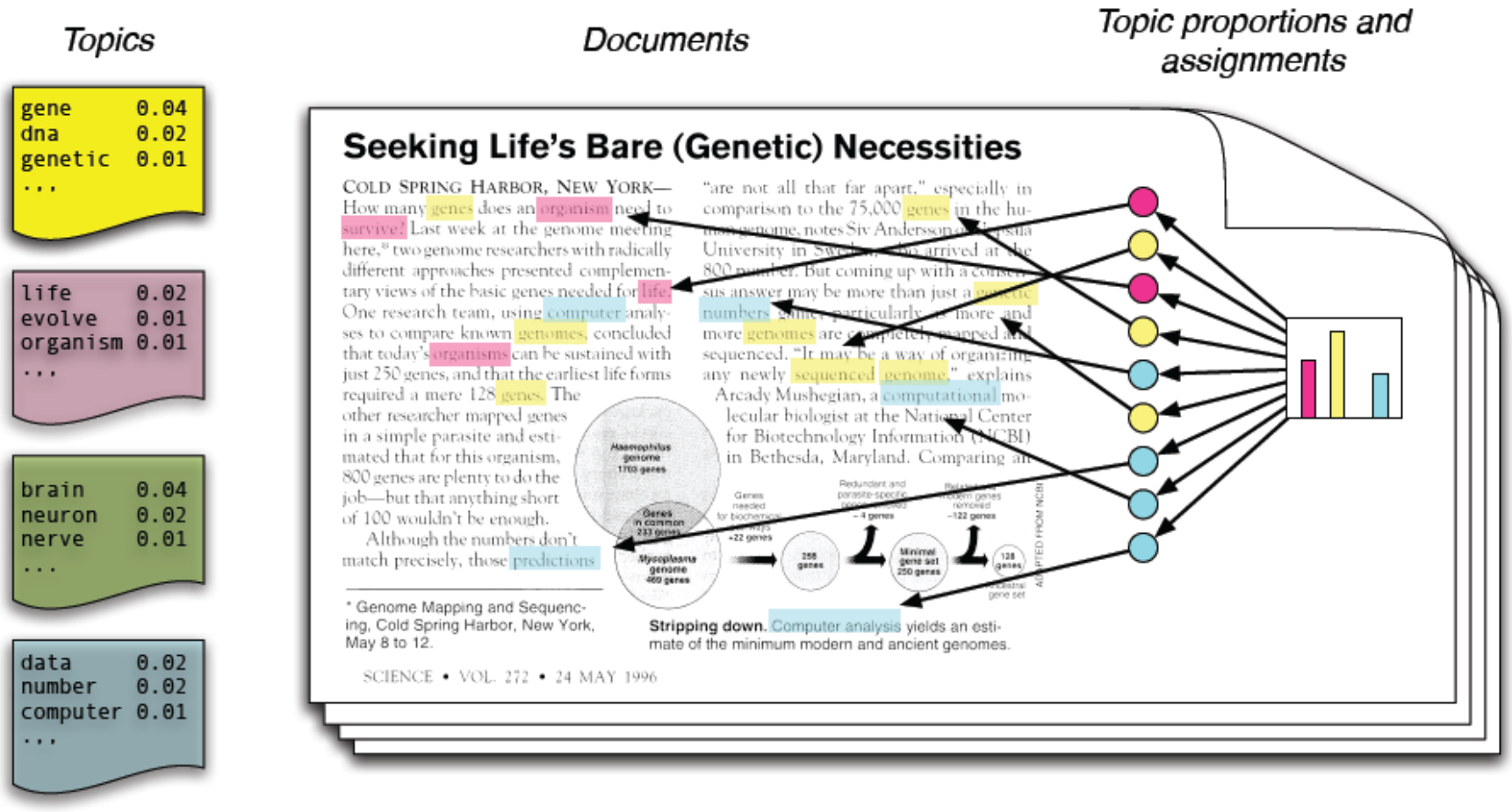
Topics are distributions over words

Documents

Documents have distribution of topics

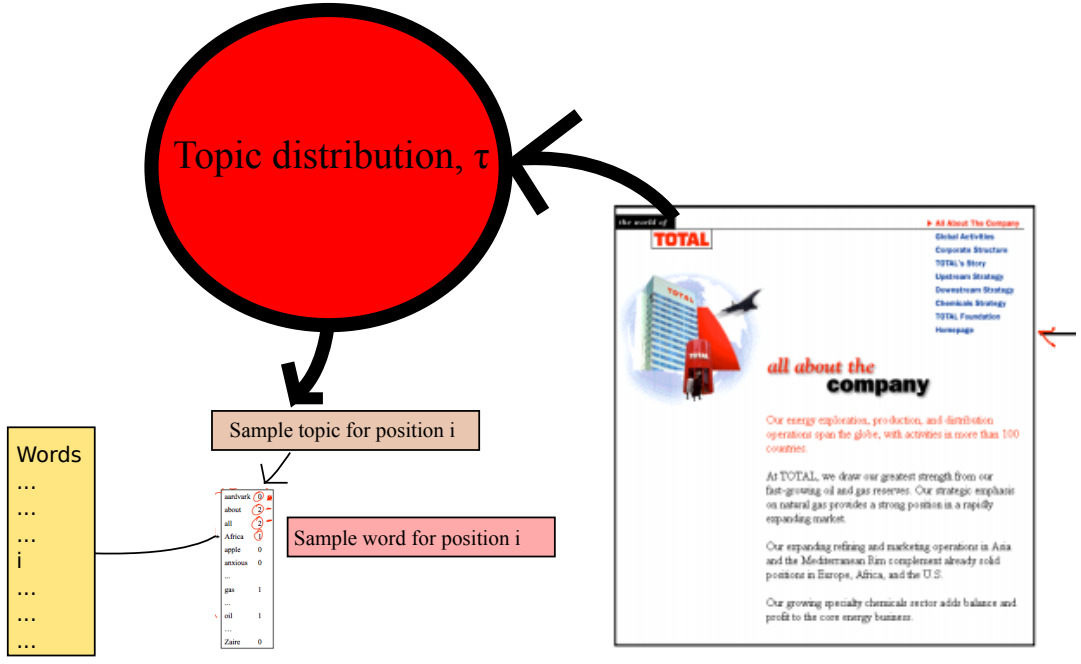


Topic Modelling



Task

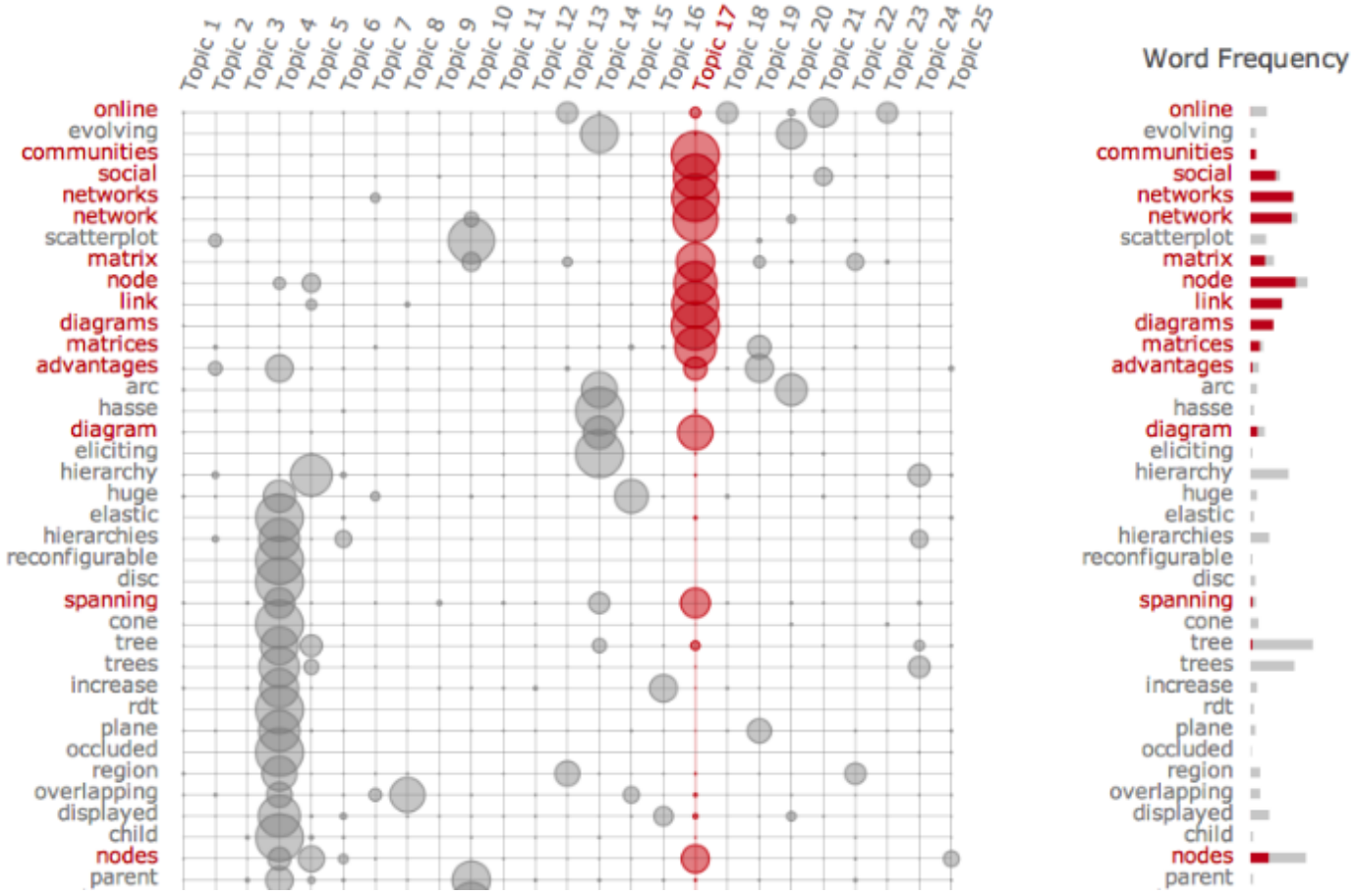
- Assume documents are generated by probabilistic model with unknown variables
- Infer hidden structure onto document
- Situate new document into model



TODO: Redo pic

Word-topic Matrix

Extracted: Word-topic matrix



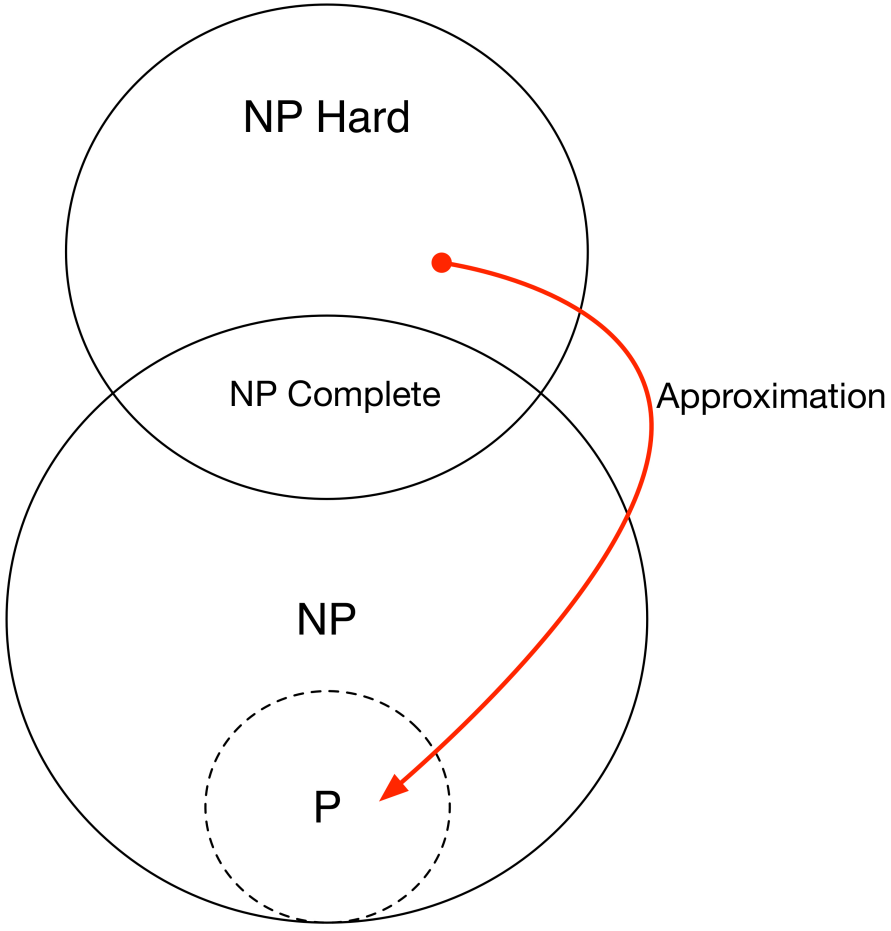
Aim: Find document-topic matrix

Anchor Words

- Word-topic distributions are separable
- There is a word unique to each topic
- Indicates document is partially about that topic
- Can learn parameters in polynomial time provided there is a large enough number of documents

Approximate Inference & Provable Guarantees

- Document-topic inference:
  - NP-hard
- Approximate techniques
- Provably polynomial-time?



Algorithm

Input: Corpus $\mathcal{D}$, Number of topics K

Output: Word-topic matrix A, topic-topic matrix R

1. Compute word-word co-occurrence matrix
2. Normalize the matrix
3. Find anchor words
4. Recover topics

Assumptions:

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

Contributions

1. Anchor Selection
   - Combinatorial rather than ILP
   - Stable in the presence of noise
   - polynomial sample complexity
2. Recovery step
   - Previous matrix-inversion approach sensitive to noise
   - Replaced with Gradient-based inference
3. Empirical comparison of algorithms

Words as vertices

TODO: Explain model of words as vertices. Lead into anchor words being words on the convex hull

Convex Hull

TODO: Explain how convex hull is computed. give time complexity for higher dimensions. Explain how convex hull "encloses" all the words within. Link aspects of problem to visual aspects

Previous method

TODO: ILP approach

New method

TODO: Iterative approach

Problem

TODO: What is topic recovery

Previous method

TODO: Overview of approach

Matrix Inversion

TODO: How matrix inversion works TODO: How this recovered topics

New method

TODO: Overview of approach

Bayes' Rule

TODO: Bayes rule and how it relates to new method

Experiments

TODO: overview of the experiments run

Metrics

TODO: Metrics

Documents

TODO: Talk about semi-synthetic documents, real documents and the need for both

Results

TODO: describe results. Iterate through each experiment, and each document type, reporting the computed metrics for each. This may need to be split up into more slides by either experiment or document type

Summary

TODO: Put the paper's conclusion into dot point form

Comments

TODO: Do we need to comment on the paper?
Are there things that we wish they had reported but didn't?
Are there things that we really liked that they reported?
Check the marking guidelines about what exactly we need here

Future Work

TODO: They didn't have a future work section but they really should have. We can make one up and maybe comment that they didn't put a future work section

Thanks!

Any questions please email either of us:

Vanush Vaswani
vvas****@uni.sydney.edu.au

Kristy Hughes
khug2372@uni.sydney.edu.au