

# 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

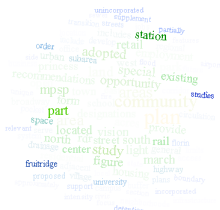
# Information Overload



# Effective Organisation



# Topics



- 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

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  
...

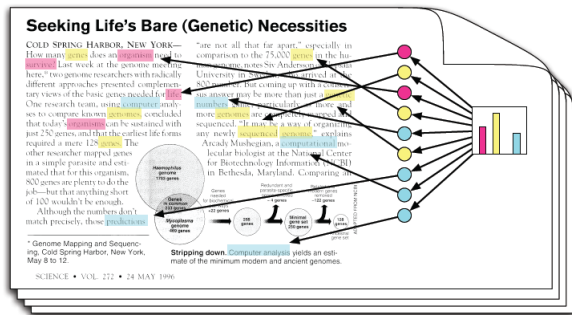
Topics are distributions over words

# Model of Documents

Documents have  
distribution of  
topics

Documents

Topic proportions and  
assignments





# Topic Modelling

## 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 Anderson, a geneticist at the University of Sydney, 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 sequenced and analyzed. "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



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

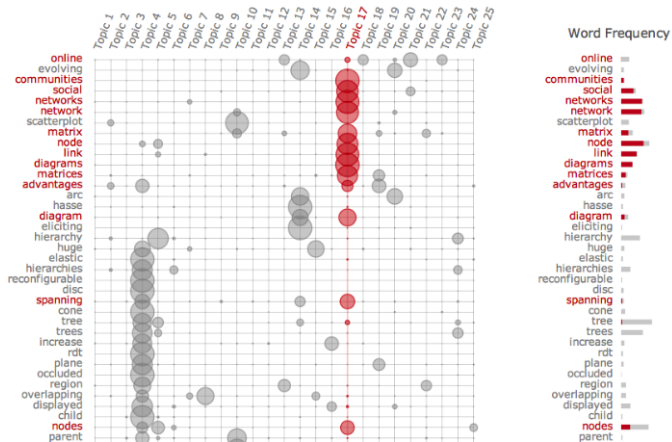
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments

# Word-topic Matrix

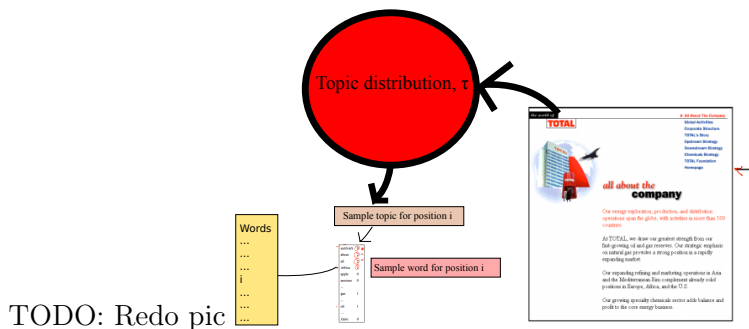
## Extracted: Word-topic matrix



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

- 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