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

Topic modeling

- Statistical modeling
- Discovers hidden thematic structure (topics) in a collection of documents
- Help to develop new ways to:
  - Search
  - Browse
  - Summarize

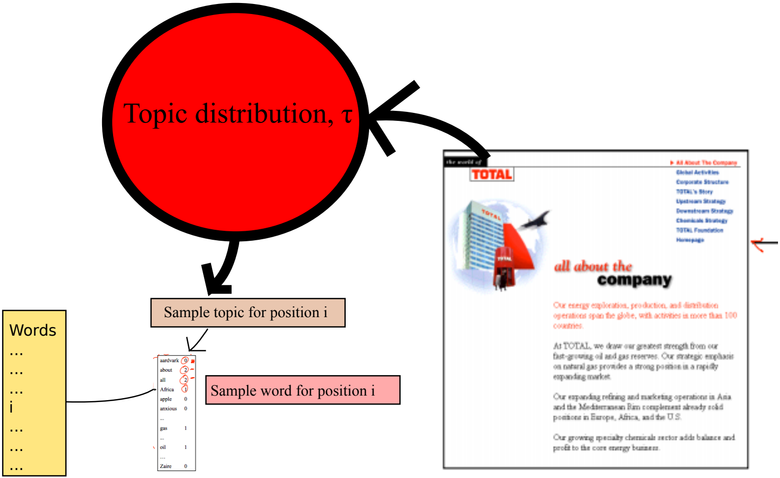
Recent Work

- Posterior inference is NP-hard (worst case)
- Approximate techniques used (SVD, Variational Inference, MCMC)
- Provably polynomial time algorithms: Statistical recovery problem
- Anandkumar et al. (2012)
  - Third-order moments
  - Assumes topics are not correlated
- Arpra et al.
  - Second-order moments
  - Assumes topics are separable
  - i.e. There exists an anchor word for every topic
  - Steps: find anchor words, reconstruct topic distributions

Contributions

- Combinatorial anchor selection algorithm
  - Assumes separability
  - Stable in presence of noise
  - Polynomial sample complexity
- Simple probabilistic interpretation of the recovery step
  - Arora et al. (2012) use matrix inversions $\rightarrow$ sensitive to noise
  - Replace matrix inversion with gradient-based inference
- Empirical comparison between recovery-based algorithms and existing likelihood-based algorithms

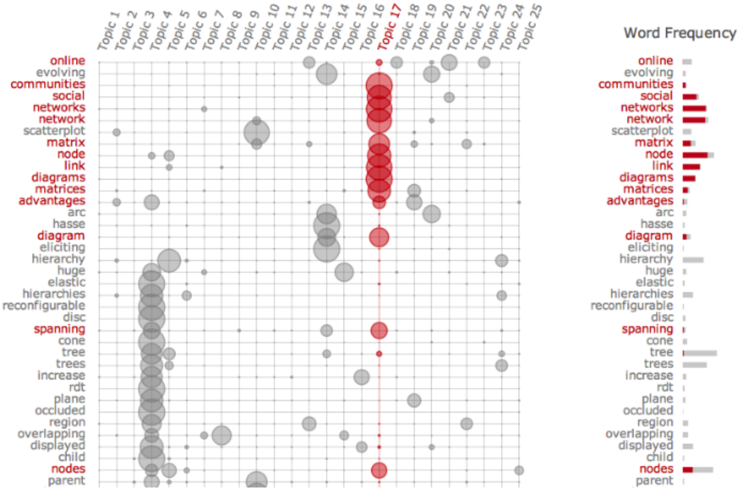
Modeling Process



Task

- Find the word-topic matrix, A
- Essentially statistical recovery
- Hyper-parameters of topic distribution

Word-topic matrix



Mathematical Prerequisites

- For each pair of words w_1 and w_2
- And their topic assignments z_1 and z_2
- The elements of the word-topic matrix are:

$$A_{i,k} = p(w_1 = i | z_1 = k)$$

- Word co-occurrences:
    - $Q \rightarrow$ joint probability of words occurring together
    - $Q_{i,j} = p(w_1 = i, w_2 = j)$
    - $\bar{Q}_{i,j} = p(w_2 = j | w_1 = i)$

Convex Hulls

For an anchor word (row) in the co-ocurrence matrix

$$Q_{s_k,j} = \sum_{k'} \underbrace{p(z_1 = k' | w_1 = s_k)}_{= 1 \text{ because of the anchor word property}} p(w_2 = j | z_1 = k')$$

$$= p(w_2 = j | z_1 = k) = C_{i,k}$$

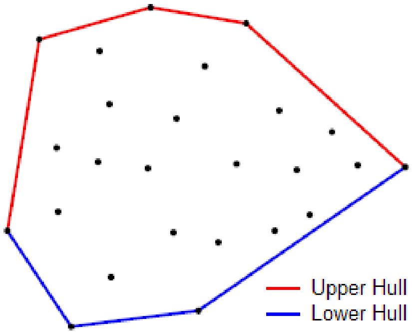
For any other row

$$\bar{Q}_{i,j} = \sum_k p(z_1 = k | w_1 = i) p(w_2 = j | z_1 = k)$$

But this is clearly a convex combination of anchor words

$$\bar{Q}_{i,j} = \sum_k C_{i,k} \bar{Q}_{s,k}$$

Convex Hulls

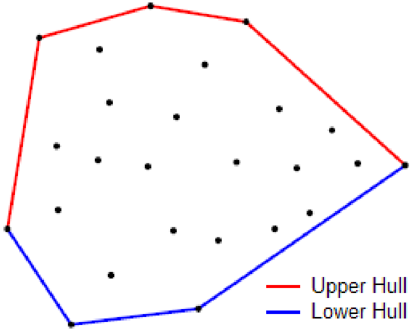


Bayes Rule

$$p(w_1 = i | z_1 = k) = \frac{p(z_1 = k | w_1 = i) p(w_1 = i)}{\sum_i p(z_1 = k | w_1 = i') p(w_1 = i')}$$

1. Introduction
2. Background
3. Topic Recovery via Bayes' Rule
4. Anchor Words
5. Experimental Results
6. Conclusion

Old algorithm



Previous algorithm:
Anchor words → Vertices on convex hull
Use ILP to find vertices of hull → Inefficient

New Algorithm

- Iterative algorithm
  - Finds farthest point from anchor words span
  - This point becomes a new anchor word
- New anchor words are most different from current anchor words
- Terminates after a set number of anchor words are found

1. Introduction
2. Background
3. Topic Recovery via Bayes' Rule
4. Anchor Words
5. Experimental Results
6. Conclusion

Methodology

Efficiency

Semi-synthetic documents

Real Documents

Conclusion