Motivation

Topic modeling facilitates a quantitative analysis of textual unstructured data. Most of today’s topic-mining algorithms model text as n-grams. Although concise, this neglects the association between non-consecutive words. Thus, our research hopes to improve the outcome of topic modeling techniques by using graph-based methods; This preserves word order and document semantics, as well as provides the audience with human-comprehensible descriptions and visualization of topics.

A topic is typically modeled as two distributions:

1. A multinomial distribution \( \phi_k \) over words

   \[ \phi_{k,x} = P(x|k) \]

   is the probability of seeing word \( x \) in topic \( k \), frequent terms related by a common theme are expected to have a large probability

2. A multinomial distribution \( \theta_{d,k} \) of topics in document \( d \)

   \[ \forall d \in D \text{ where } D \text{ is the document collection} \]

   \[ \theta_{d,k} = P(k|d) \]

   is the probability of seeing topic \( k \) in doc. \( d \)

Through a bit of theory and experimentation, we developed models that produce accurate, relevant, and complete phrases that leverage word order and topic distributions

Detailed Overall Approach

1. Use Latent Dirichlet Allocation (LDA) [3], to create the term \( \phi_k \) and document \( \theta_{d,k} \) distributions

2. Represent text collection as simple directed weighted graph for each topic

   \- Each word \( x \) is a node with weight \( \phi_{k,x} = P(x|k) \)
   \- Word order and co-occurrence defines the edges
   \- Edges are weighted based on \( \theta_{d,k} = P(k|d) \)

3. Mine these graphs for phrases.

   \- Sort nodes based on \( \phi_{k,x} \)
   \- Pick node with highest \( P(x|k) \)
   \- Traverse to the most promising nodes based on connecting edge weights
   \- Terminate if \( \prod_{x \in \text{path}} P(x|k) > \text{threshold} \) or \( \beta \) outlinks

An Example

This schematic example represents a simplified version of how a directed weighted graph is built using a corpus of documents related to the abduction of Chibok schoolgirls by Boko Haram [2]. The edges are blue if the relation between two nodes in the corpus is highly likely and gray otherwise.

Future Directions

We plan to:

- Experiment with different traversal methods to refine results
- Develop a method that yields the optimum threshold value
- Evaluate quantitatively the accuracy of the algorithm.
- Leverage graph similarity & subgraph matching methods [1] to compare topics

Our source-code has already been open-sourced [4]

References


Acknowledgments

We would like to thank Arpit Khanuja for his continued contributions to this project as well as Professor ChengXiang Zhai and the PURE Committee for their support of our work.