Topic Modeling

A 10-slide primer

David Newman

newman@uci.edu

What is the topic model?

- The topic model is an algorithm that automatically learns topics (themes) from a collection of documents
  - It works by observing words that tend to co-appear in documents, for example *gene* and *sequencing*, *or climate* and *change*
  - The topic model assumes each document exhibits multiple topics
  - The topic model learns topics *directly from the text*

Why is it useful?

- Topics can help organize, search and understand a document collection. One can ask:
  - What topics are in this document?
  - What documents are related to this topic?

What is a topic?

- The topic model learns a set of topics
- Each topic is displayed by showing its top-20 words, for example:
  - dark_matter cosmological cosmology universe dark_energy lensing survey CMB redshift cosmic mass galaxy scale galaxies gravitational measurement power_spectrum parameter observation structure ...
  - This is a topic about *Dark Matter, Dark Energy and Cosmology*
  - Note: The ellipsis indicates that the list of words continues

Example of topics in an award

0910908: Intelligent Tracking Systems that Reason about Group Behavior

Top-4 topics in this award:

(topic 400) Computer Vision
(topic 512) Small Mammals
(topic 319) Tracking Systems
(topics 94) Complex Systems

The ability to reason about the complexity of living organisms in diverse environments is one of the hallmarks of intelligence. In this project the PI and her interdisciplinary team of investigators will design computer vision algorithms for intelligent tracking of large groups of living individuals in three-dimensional space. She will develop specific systems for tracking groups of microorganisms, bats, birds, and humans. And she will formulate machine learning methods for analyzing group behavior, specifically the conditions for formation and dispersal of groups, and the interactions of individuals within a group. An important innovative aspect of this research is the systematic and comprehensive approach to reasoning about the motion of large groups of living organisms observed in video data, independently of whether they happen to be humans, animals, or cells ...

Two modes of usage

- **Learning**
  - Given a set of documents, learn set of topics
  - We've done this for NSF. Topics are here: http://www.ics.uci.edu/~newman/nsf/20120105/topics.txt
- **Inference**
  - Given one document (e.g. a new proposal), and set of topics, infer topics in this document
  - We're doing this for NSF submitted proposals on an ongoing basis
  - Store the top-4 topics for each proposal

Use topics to characterize things

- One proposal or award
- A set of proposals (in a program)
- An NSF division
- An investigator
- A reviewer
- An institution or center
- … etc
- *And track changes over time*

Use topic model to answer these types of questions

- Who are researchers who are qualified to review this proposal?
- How (topically) have the grants in SciSIP changed between 2007 and 2009?
- What grants, for which I am not the program officer, are similar to my set of grants?
- My program is Ion Channels. What topics are addressed by grants not in my program, but do mention Ion Channels?
- What is the topical makeup of grants in Statistics that don't mention "Bayesian"?
- What grants are related to Climate Change?
- How much does NSF spend on nanotechnology and nanomaterials?
- What topics do program (or program manager) X and Y have in common?
- Are topics associated with EAGER and RAPID funding different from those with standard funding?

FAQ

- Q: Do topics come with labels?
  - A: No, but it is easy to have subject matter experts label topics. We have done this for our 1000 NSF topics
- Q: Can topics be edited?
  - A: Yes. Sometimes topics have unfitting words, and sometimes topics are inappropriately assigned to documents. This can be manually edited and corrected
- Q: How to set number of topics?
  - A: The appropriate number of topics depends on the application. For NSF, we decided that 1000 topics would provide good resolution of categories of research

References

- The topic model is about 10 years old, and has been researched extensively over the last decade
- It evolved out of Latent Semantic Indexing
- References (very technical)
  - Blei, Ng, Jordan (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research
  - Griffiths & Steyvers (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences
  - A nice tutorial: http://www.cs.princeton.edu/~blei/kdd-tutorial.pdf
- The topic model is also known as *Latent Dirichlet Allocation*
- Please email me with any questions: newman@uci.edu