Understanding and Using Topic Modeling

Using inferred document clusters

Philip Robinson

Presented to Knowlege Mavens

February 2, 2019

Jet Propulsion Laboratory
California Institute of Technology

Presentation Overview

What is Topic Modeling

Our goal: mathematically model topics from a corpus

Topic modeling is a text processing technique for automatically grouping documents by topics. This is usually used as a strategy to describe documents in low dimensional space or an exploratory tool for document collections.

Examples

In practice, this requires many more documents

The Tourist huddles in the station While slowly night gives way to dawn ; He finds a certain fascination In knowing all the trains are gone.

The Governess up in the attic Attempts to make a cup of tea ; Her mind grows daily more erratic From cold and hunger and ennui.

The Journalist surveys the slaughter, The best in years without a doubt; He pours himself a gin and water and wonders how it came about.

- Food
- Travel
- Time

From this annotation we know that Document 2 and 3 are about Food and Time

What can I solve?

- Find similar document pairs
- Cluster documents into groups with similar content
- Find relevant documents to a query or user's interests
- Explore shape of document collection

Topic modeling can usually be extended to address many other problems, and document embeddings can be used to inform downstream models.

Applications of Topic Modeling

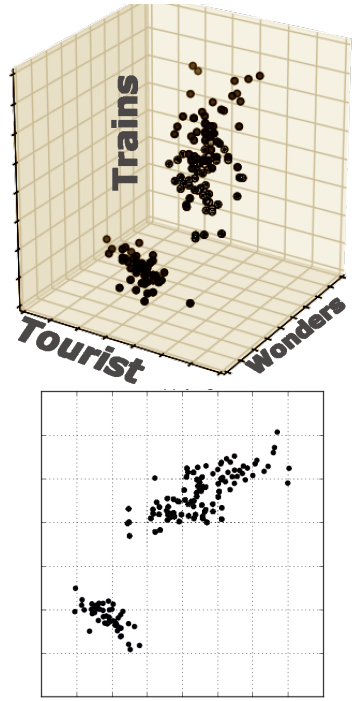

Figure: Dimmentionality Reduction

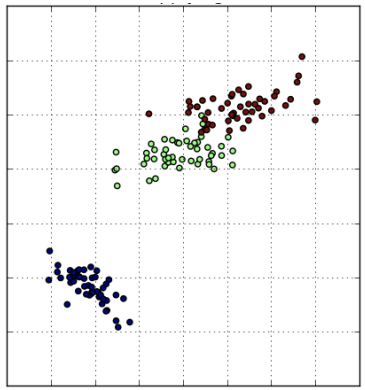

Figure: Cluster Points

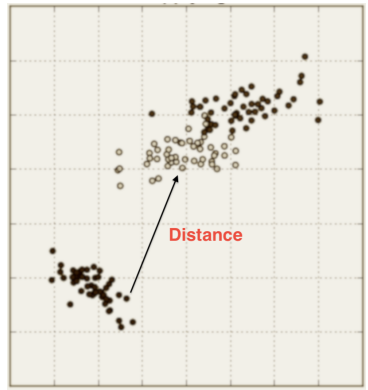

Figure: Exploratory Data Analysis

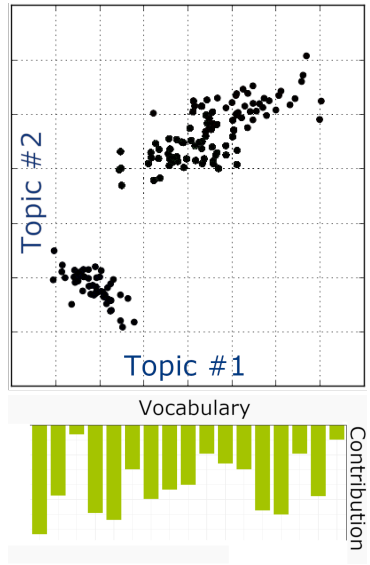

Figure: Analysis of Topics

Goals of generative models

A generative model

- Assume/Generalize how data could have been generated
- Fit distributions that describe the generalization
- Ask questions about the generalization in relation to data
- Ask questions about data in relation to the generalization

Generative models are much easier to extend, because they abstract the model from it's linear algebra dependencies.

Topic modeling generalizes how a document is generated by claiming that words come from topics, and documents have multiple topics.[1]

[1] this is not a language model

Latend Dirichlet Allocation

Bayesian extension to PLSA

- Represent document as Bag-of-Words[2]
- Model/Fit topics as mixture of words
- Documents are projected into or sampled from topic-space-distribution

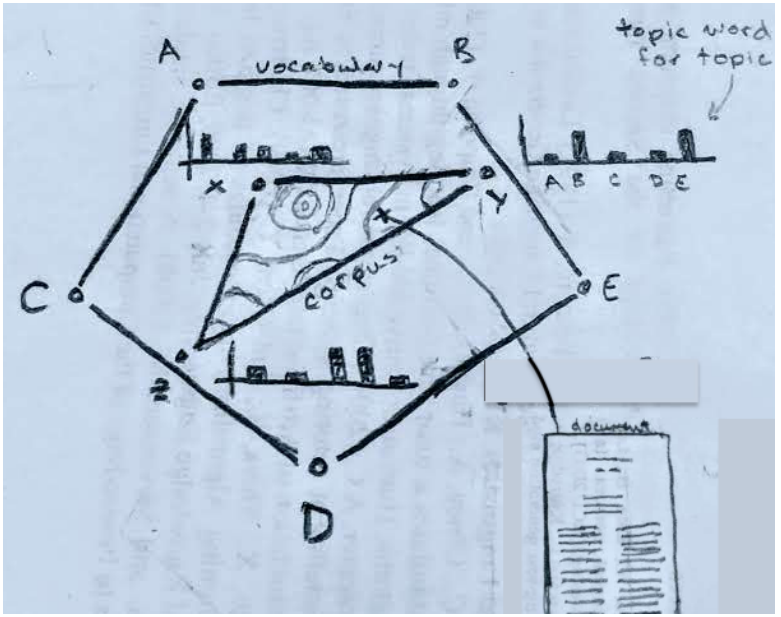

Figure: Latent Dirichlet Allocation

Enormous body of work extending this model to address more specific problems.

[2] equivalent to multinomial over the vocabulary

Looking at top words

Mitigating apophenia is hard, topics difficult to interpret

Topic #1
- server
- connected
- access
- workstation
- outage
- user

Topic #2
- mode
- instrument
- safe
- spacecraft
- anomaly
- recovery

Topic #3
- uplink
- station
- dsn
- spacecraft
- lock
- ace

Although the model better describes our generation process, from the perspective of topics, it can be difficult to know what these topics actually represent.[3] *This may require experts who are immune to apophenia.*

[3] Supervised LDA attempts to addresses this concern, also applies to sentiment analysis

Extensions

- Entity Boosted Topic Model (ETM)
- → Author Topic Model (ATM)
- Hierarchical Dirichlet Process (HDP)