Topic Modeling

Prof. Eric A. Suess

April 14, 2021

Topic Modeling

Unsupervised Learning

Clustering of documents

Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for.

Latent Dirichlet allocation (LDA)

- It treats each document as a mixture of topics, and each topic as a mixture of words.

This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

LDA

- Every document is a mixture of topics.
- Every topic is a mixture of words.

Example

The example in the book runs topic analysis on the Associated Press articles from around 1988.

The two topics found are *Financial News* and *Politics*.

Document-topic probabilities

- gamma