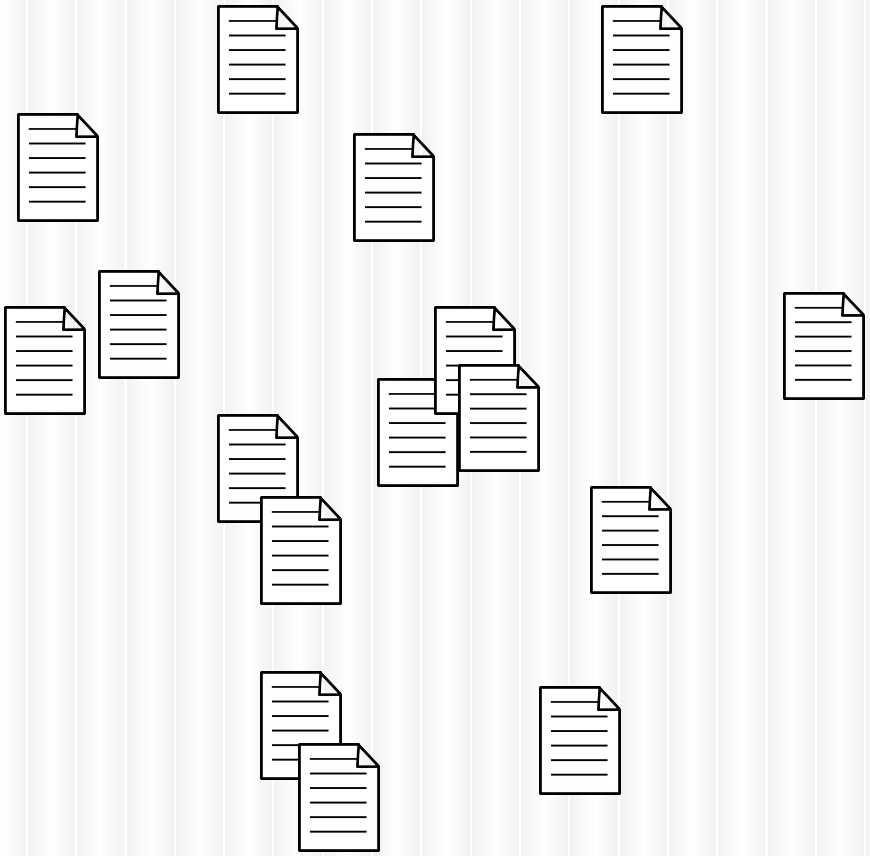
Latent Dirichlet Allocation

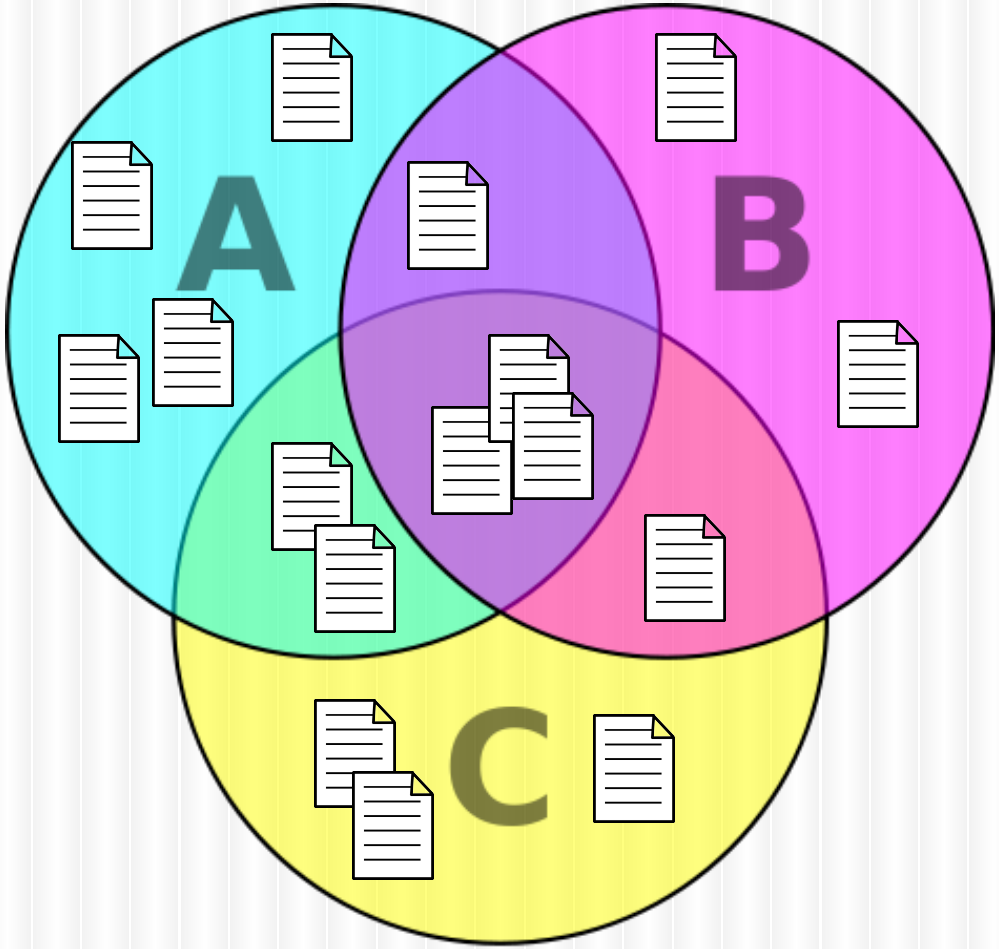
It Sounds Complicated!

Cameron Carpenter
UMBC CMSC 676
April 28, 2015

Topic Modeling



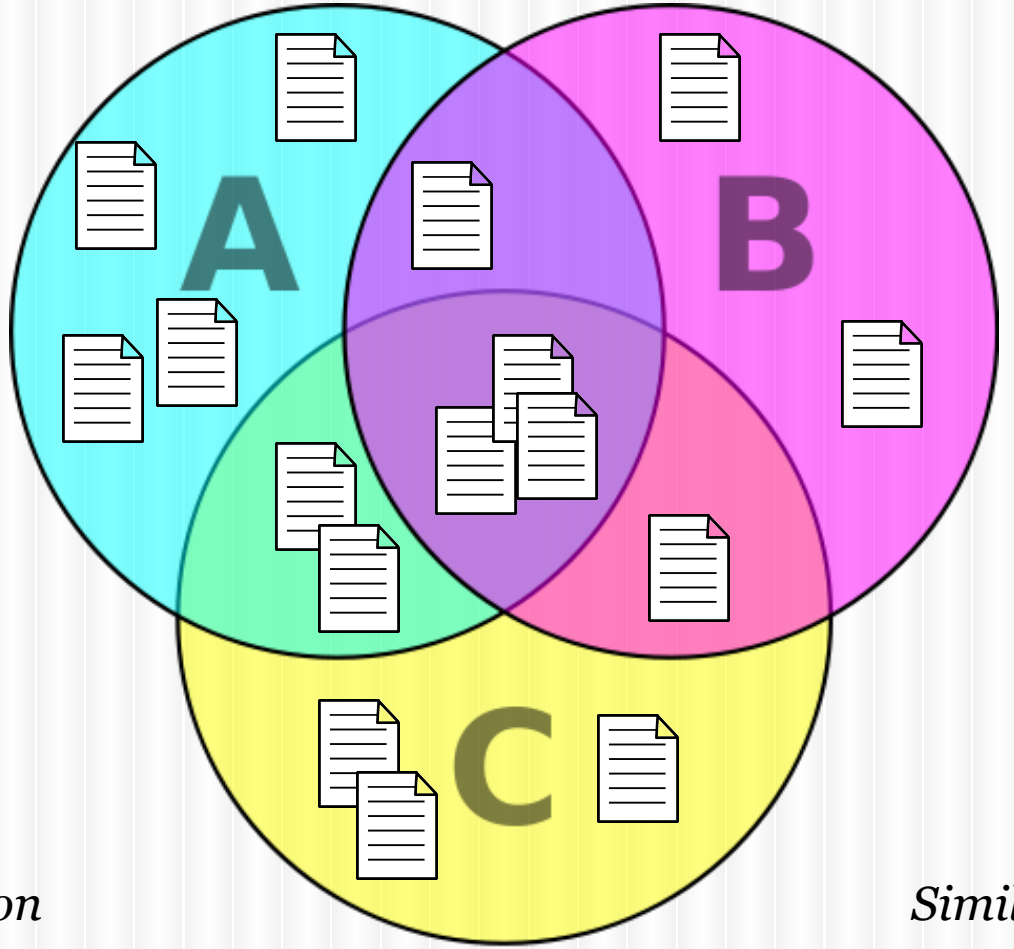
Topic Modeling



Topic Modeling

Classification

Summarization

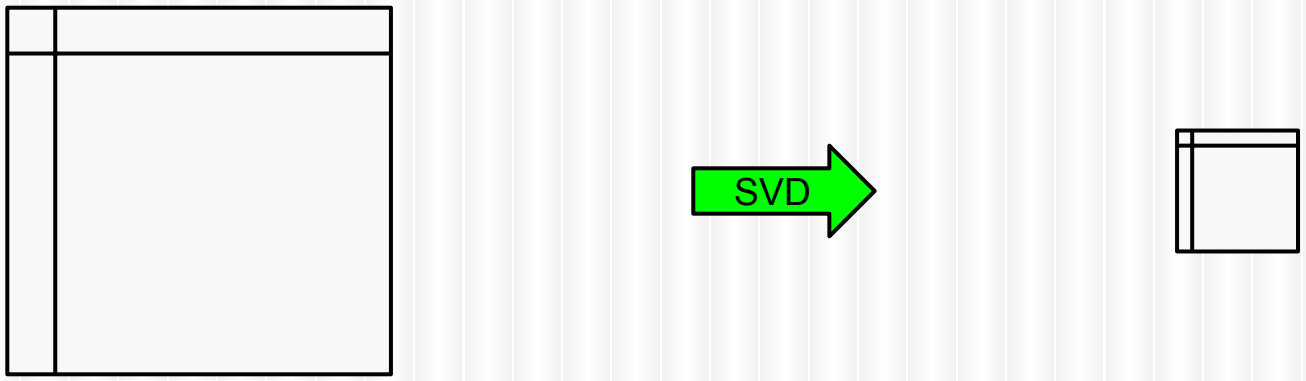


Novelty Detection

Similarity

Latent Semantic Indexing

Find a linear subspace of features that captures most of the variance



...But how do you evaluate it?

pLSI (aspect model)

Every document is a mixture of topic proportions

Key Idea

Given a generative model of text, one can fit the model to data using statistical methods.

This Isn’t a Grant Proposal

Let me try that again…

Key Idea

If you make a few assumptions about how the documents were created, you can use math to guess what the topics might be.

This is LDA (really!)

1. Make a list of all relevant topics
2. Until you have enough documents:
    a. Make a new empty document
    b. Pick a short list of topics for the document to be about*
    c. Until the document has enough words in it:
        i. Pick a topic from the short list
        ii. Pick a word from that topic**
        iii. Write that word down

Topics

probability distributions over words

Animals		**Food**		**Music**	
Pig	0.00053	Bread	0.00159	Note	0.00243
Lemur	0.00013	Sushi	0.00029	Bass	0.00198
Bass	0.00009	Table	0.00143	Jam	0.00089
Cat	0.00102	Jam	0.00073	Violin	0.00252
...		...		...	

An Example

- Document 1
  - 60% Food, 40% Music
  - note bread sushi table bass note jam jam table bread
- Document 2
  - 20% Animals, 80% Food
  - bread table table bread pig sushi cat bread sushi table

The Real World

- You don't know...
  - ...how much of each topic a document contains
  - ...which topic a word in a document belongs to
  - ...what words get what weight in each topic

*All of these attributes are hidden, or **latent***

The Posterior

Given that you only know the words in the documents, how can you figure out the rest?

Sampling Methods

Variational Methods

Sources

Bishop, Chris. Pattern recognition and machine learning, volume 4. 2006.

Blei, David. Introduction to probabilistic topic models. Communications of the ACM, 2011.

Blei, David, et. al. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993-1022, 2003.

Reed, Colorado. Latent Dirichlet Allocation: Towards a Deeper Understanding. University of Iowa, 2012.