Topic models are a representation of probability.

LDA: latent dirichlet

A single document [on what probability] covers a specific topic (word or phrases).

Q: no dictionary are needed before hand.

What is a topic model good for?

* Indexing;
* Summarization
* Discovery

Topic models are more for pattern matching, rather than semantics reading.

What kind of data do you have?

* Structured versus unstructured
  + Did human write it? Is it in xml format?
* Latent versus categorical analysis
  + Do you know what you are looking for in advance?
* Big data
  + Will humans code it too?
  + Is there enough for lda to bite into?
* LDA shines with large amounts of unstructured data that you need to analyze latently.

Issues to be aware of

* Overfitting
* Overreliance
* Too much distance from the text

For each topic, actually, it is a group of separate words of a topic.

Libguide@uiuc

The speed depends on corpus.

Software Mallet

http://guides.library.illinois.edu/c.php?g=750453&p=5464739