Topic Models

Dr. Travis G. Coan
T.Coan@exeter.ac.uk

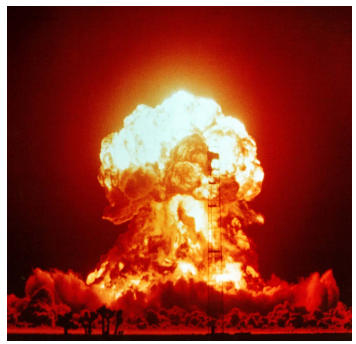
What's a **topic model**?

Generally speaking, **topic models** offer an **unsupervised** approach to extracting the themes (or "topics") present in a large corpus of data. Topic models are described in various ways across the literature:

1. Yet another set of clustering algorithms in a long line of clustering algorithms.
2. A data reduction technique
3. An unsupervised classification algorithm
4. (my take) A set of algorithms to automatically (sort of) learn the dictionary keywords described last week.

Topic models are all of these things!

The evolution of **topic models** in the literature



Latent Dirichlet Allocation (LDA)
Blei et al 2003

1988: Latent Semantic Analysis (LSA)
[or Latent Semantic Indexing (LSI)]
Deerwester et al. 1988

Probabilistic Latent Semantic Analysis (PLSA)
Hoffman 1999

Time

Latent Dirichlet Allocation (LDA)

The LDA model

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

*(**?**, p. 996)*

The LDA model

Latent Dirichlet allocation (LDA) is a **generative** *probabilistic model of a corpus. The basic idea is that documents are represented as* **random mixtures** *over* **latent topics**, *where each topic is characterized by a distribution over words.*

*(**?**, p. 996)*

The LDA model

LDA provides a statistical framework for understanding the latent topics or themes running through a corpus by explicitly modelling the random process responsible for producing a document, assuming that each document is made up of a mixture of topics, as well as a mixture of words associated with each topic. For instance, the document you are reading at this moment includes a mixture of themes such as "climate scepticism" and "text analysis," and these themes tend to use different language—the topic "climate scepticism" is likely associated with the word "denial," whereas the topic "text analysis" is associated with the word "random." Moreover, this process is probabilistic in the sense that we could have used the term "stochastic" instead of "random" in the previous sentence.

(Boussalis and Coan, 2016, p. 92)

Generative models

Bayesian models—whether of text or any other data structure—are said to be **generative**. That is, they define a joint probability distribution over all random variables in a model, both observed and hidden.

Intuitively, the **generative process** is "the imaginary random process by which the model assumes the documents arose" (**?**, p. 78). Put simply, it is a "story" for how documents came to be.

The LDA's generative story

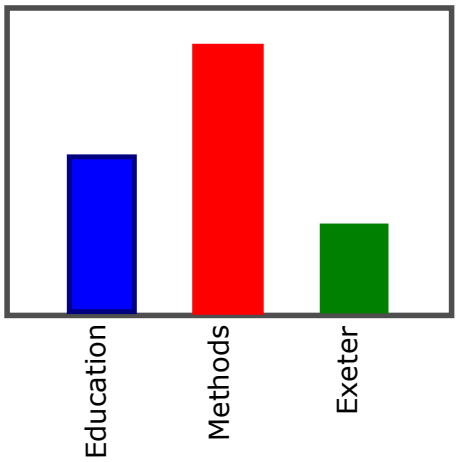
The **generative story** for the following sentence:

"Social science students should know text analysis."

?

How do we generate word 1?

Topic distribution (theta): Topic propotions are assinged using a Dirchlet distribution



Step 1: Sample a topic one of *k* topics using a multinomial distribution

Topic = Education

Step 2: Given the topic from Step 1, sample a word using a multinomial distribution

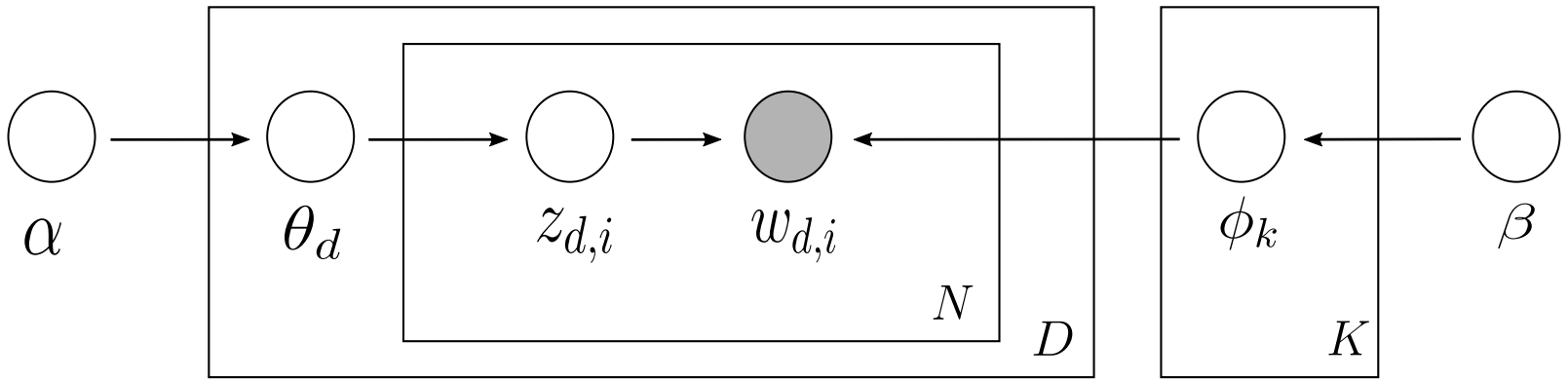
Word = social

Move to next word and repeat

Word distribution (phi)

Education	student (.20), class (.18), social (.08)
Methods	stats (.14), text (.13), data(.05)
Exeter	exeter (.23), cathedral (.10), uni (.09)

The LDA's generative story (plate notation)



The model parameters are as follows (from left to right):

α: prior distribution for the *topic distribution* (1 x K).

θ_d: the *topic distribution* for document d (1 x K)

$z_{i,n}$: topic assignment for word n in document d.

$w_{i,n}$: word n in document d.

ϕ_k: the so-called *word distribution*.

β: prior distribution for the so-called *word distribution*.

LDA in **Python**

Go to notebook!