Topic modeling: How to synthesize the information in a large collections of documents?

Basic approach: some clustering algorithm to find groups of similar documents.

Alternative: Topic modeling: find some latent topics of themes that are linked between the documents and words

LSA—> pLSA —> LDA

Preprocessing: 1. remove all non alphanumerical characters. 2. find all words existing in the corpus, which defines our **Vacabulary,**

Topic modeling: 3. Create Document-Term-Matrix (DTM) 4. How to make matrices smaller????

——LSA Singular value decomposition (SVD) 1000\*1000 -> 5\*1000 +5 + 5\*10000

—— select top 2 topics (higher)

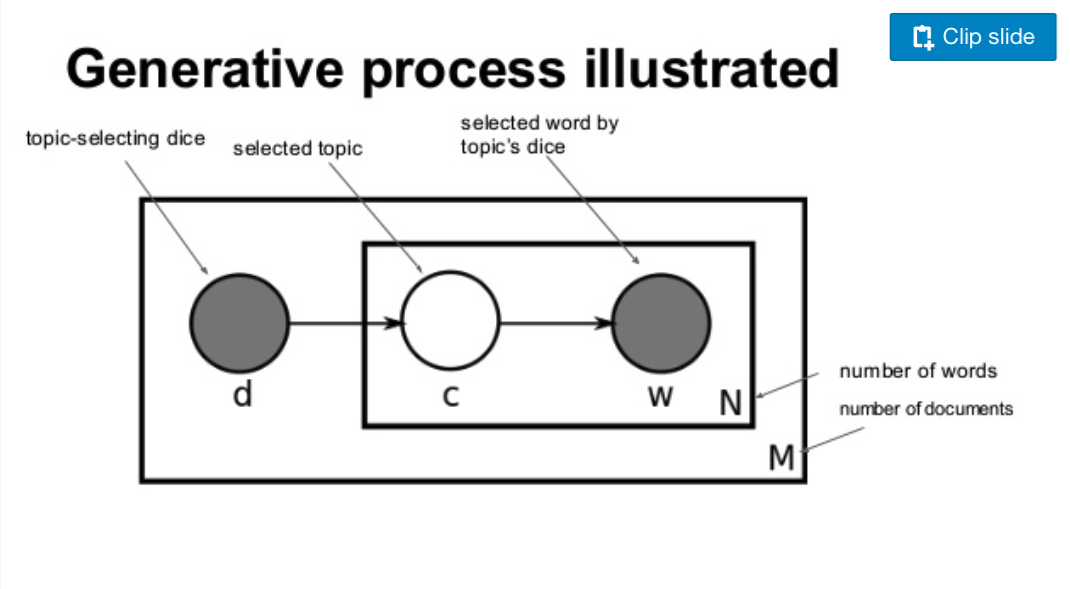
—— Essentially low-rank approximation of document-term matrix.

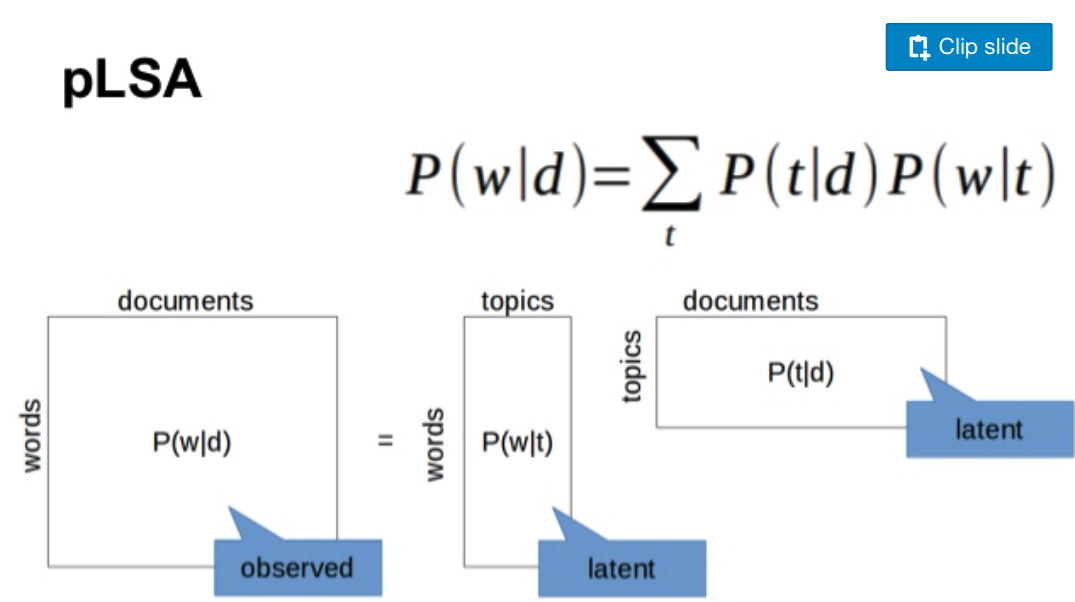
——pLSA Find a mixture of word-topic & topic-document distributions that are most likely given the observed documents.

— generative process in topic modeling

— topic-selecting dice-> selected topic —> selected word by topic's dice

d c w





？？？Overfitting? Different associations (only strongly to couple/evenly distributed)

LDA

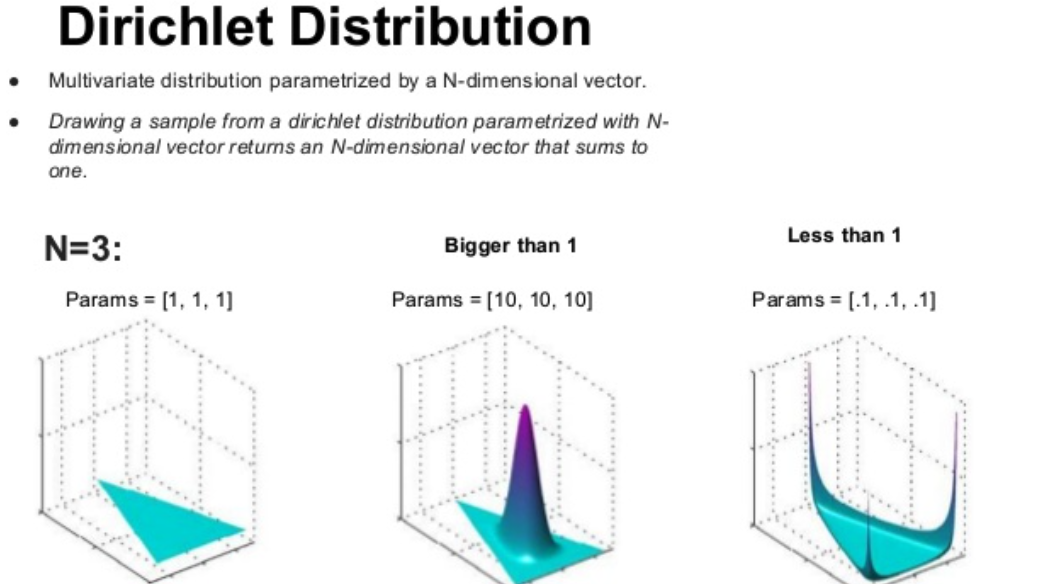
alfa K dimensional vector that defines how K topics are distributed across documents

beta V dimensional vector that defines how V words are associated across topics.

smaller values favor fewer strongly associated

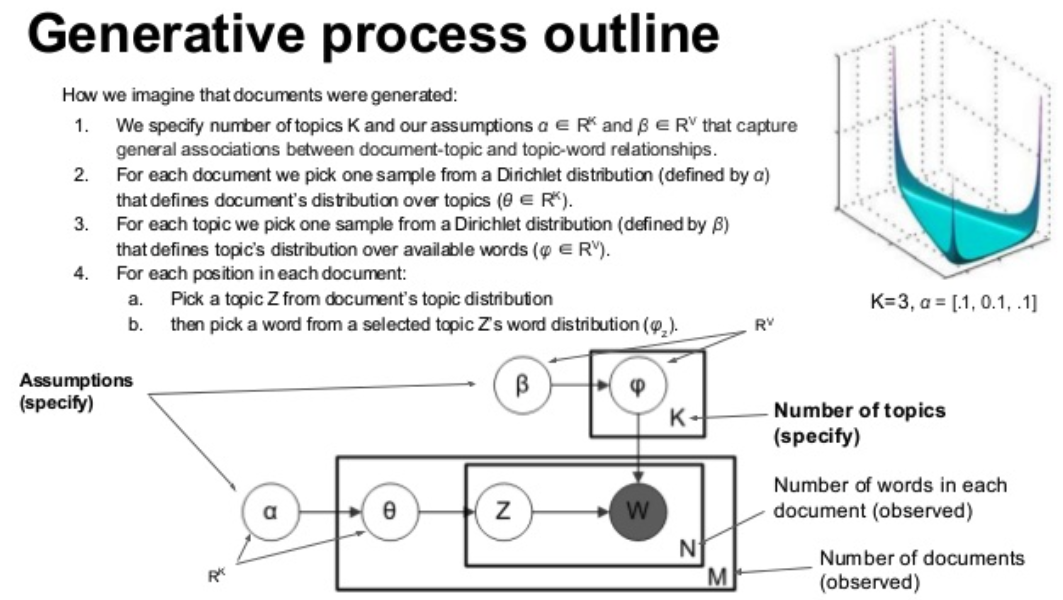
\*Typically all elements in a&b are same due to uninformative prior.

Dirichlet distribution: Multivariate distribution parametrized by a N-dimensional vector that sums to 1

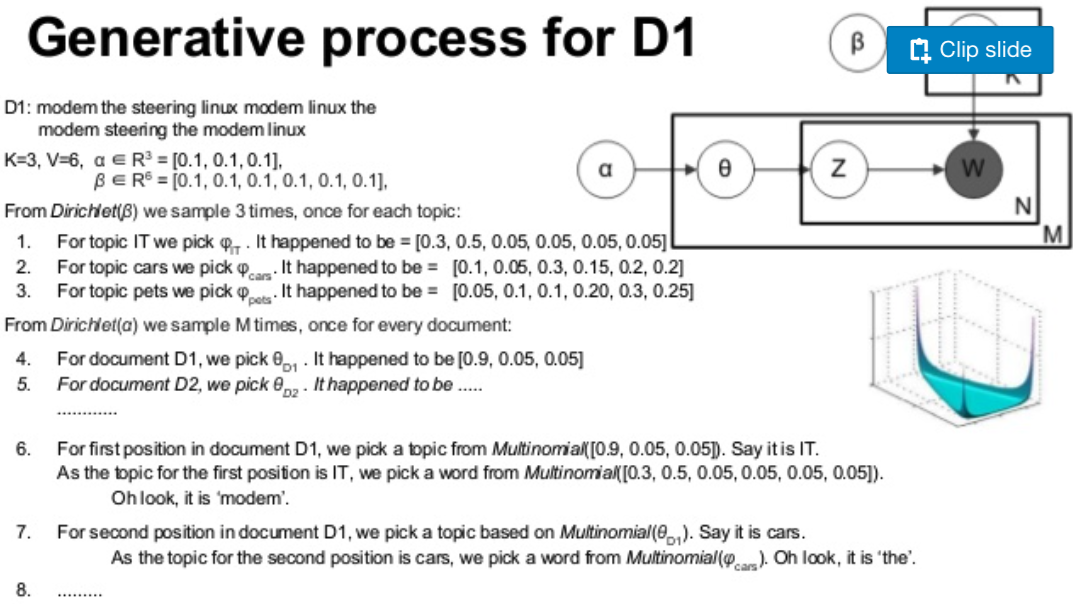


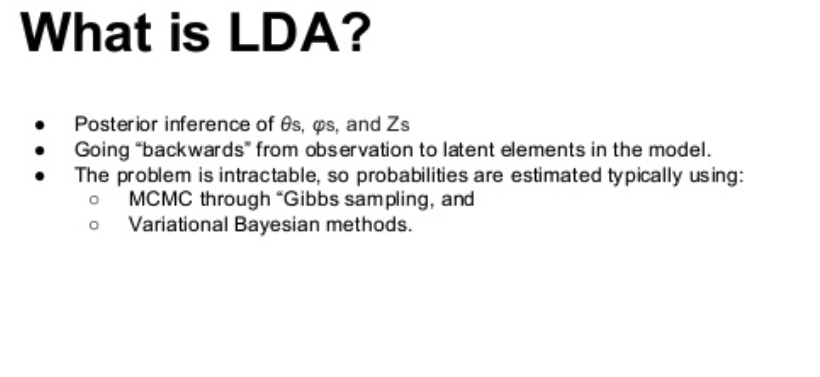
* + - How topics are distributed across docs and how words are distributed across topics are drawn from Dirichlet Distribution

Generative process



D1: modern the steering linux modem linux the modern steering the modem linux





CTM

http://www.slideshare.net/vitomirkovanovic/topic-modeling-for-learning-analytics-researchers-lak15-tutorial