Day 9: Topic Models

Kenneth Benoit

Essex Summer School 2014

July 31, 2014

Latent Dirichlet Allocation

- LDA provides a generative model that describes how the documents in a dataset were created
- Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics
- So the process is, roughly:
  1. Choose a number of topics
  2. Choose a distribution of topics, and create a document from this distribution
  3. For each topic, generate words according to a distribution specific to that topic
- The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these

Uses and applications

- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents
- Can be used to organize the collection according to the discovered themes
- Topic modeling algorithms can be applied to massive collections of documents
- Topic modeling algorithms can be adapted to many kinds of data. among other applications, they have been used to find patterns in genetic data, images, and social networks

Latent Dirichlet Allocation: Details

- Consider we have D (I) documents conisting of V (J) words each
- Assume we know that there are K topics in our corpus. Each topic describes a *multinomial* distribution over the V words, where β_k is the multinomial distribution over the words for topic k.

Latent Dirichlet Allocation: Details

- For each document, the LDA generative process is:
  1. randomly choose a distribution over topics (a multinomial of length K)
  2. for each word in the document
     - 2.1 Probabilistically draw one of the K topics from the distribution over topics obtained in (a), say topic β_k (each document contains topics in different proportions)
     - 2.2 Probabilistically draw one of the V words from β_k (each individual word in the document is drawn from one of the K topics in proportion to the document's distribution over topics as determined in previous step)
- The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these, *assuming this generative process*

More formal description of LDA

For each document:

1. draw a topic distribution, $\theta_d \sim \mathrm{Dir}(\alpha)$, where $\mathrm{Dir}(\cdot)$ is a draw from a uniform Dirichlet distribution with scaling parameter α
2. for each word in the document:
   - 2.1 Draw a specific topic $z_{d,n} \sim \mathrm{multi}(\theta_d)$ where $\mathrm{multi}(\cdot)$ is a multinomial
   - 2.2 Draw a word $w_{d,n} \sim \beta_{z_{d,n}}$

Illustration of the LDA generative process

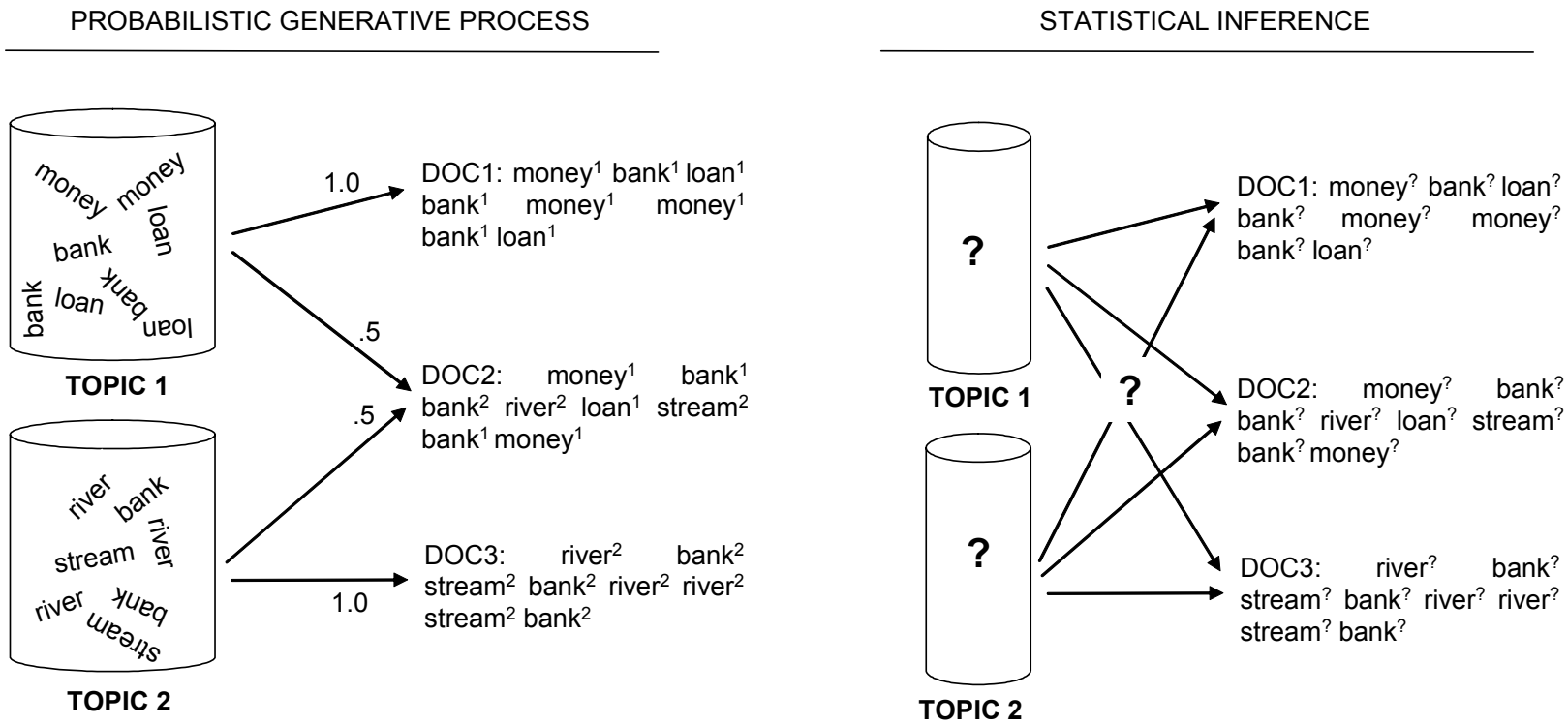


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)

Topics example

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

(from Steyvers and Griffiths 2007)

Often K is quite large!

Probablistic Topic Models

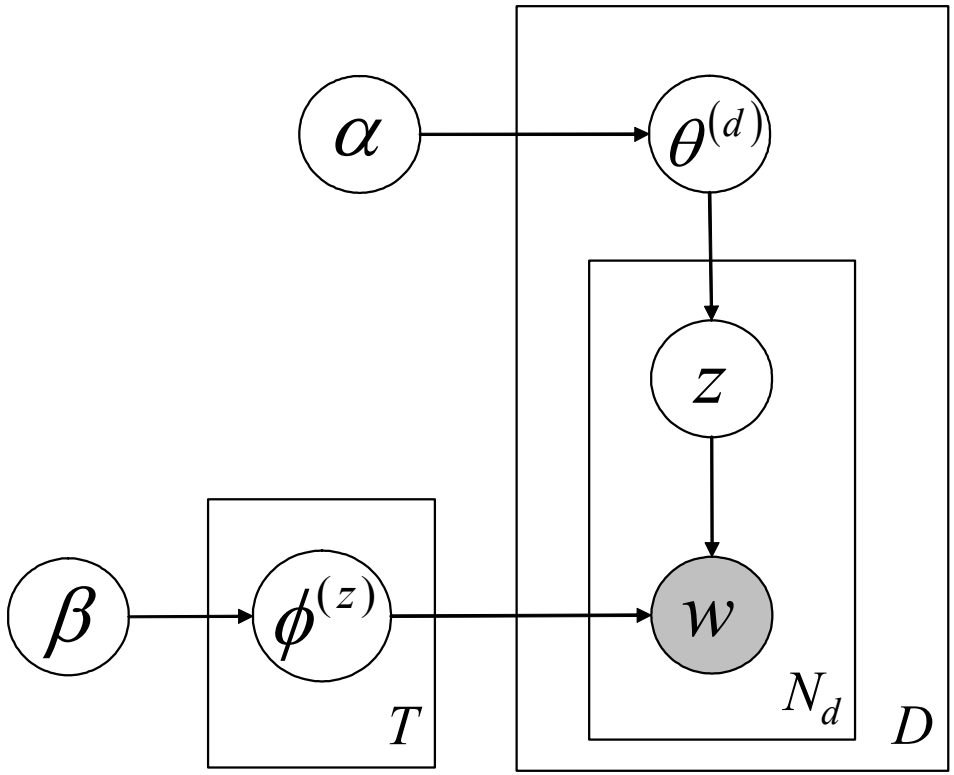
- Consider $P(z)$ as the distribution over topics z in a particular document and and $P(w|z)$ as the probability distribution over words w given topic z.
- Each word w_i in a document (where the index refers to the ith word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution
- $P(z_i = j)$ is the probability that the kth topic was sampled for the ith word token
- $P(w_i|z_i = k)$ is the probability of word w_i under topic k
- K is the number of topics
- Then

$$P(w_i) = \sum_{k=1}^{K} P(w_i|z_i = k)P(z_i = k)$$

Elements of the model

- Let $\phi^{(k)} = P(w|z = k)$ refer to the multinomial distribution over words for topic k
- $\theta^{(d)} = P(z)$ refers to the multinomial distribution over topics for document d
- Text collection consists of D documents and each document d consists of N_d word tokens, with $N = \sum N_d$ (the total tokens in the document collection)
- ϕ which words are important for which topic
- θ which topics are important for a particular document
- z_i is the assignment of word tokens to topics

Graphical model for LDA using plate notation



Interpreting a plate model:

- shaded variables are observed, unshaded are latent
- the hyperparameters α and β are treated as constants in the model (these are parameters of the Dirichlet distribution)
- arrows indicate conditional dependencies between variables
- plates (boxes) are repetitions of sampling steps, with the lower right corner variable indicating the number of samples

Estimation and the "Dirichlet" part

- The Dirichlet is the conjugate prior distribution for the multinomial, and is used in the Bayesian inference required to estimate these parameters
- Estimation is performed using (collapsed) Gibbs sampling and/or variational Expectation-Maximization
- (for us) Implemented in the `lda` library and can be used with `quanteda` dfm objects

Challenges in applying LDA

- How many topics (K)?
- How to interpret (label) the topics?
- Should we expect all topics to make sense?
- Some models are complicated and expensive to estimate