Seminar 4
Probabilistic Topic Model

Mikhail Kamrotov

Data Analysis in Politics and Journalism

Winter/Spring 2019

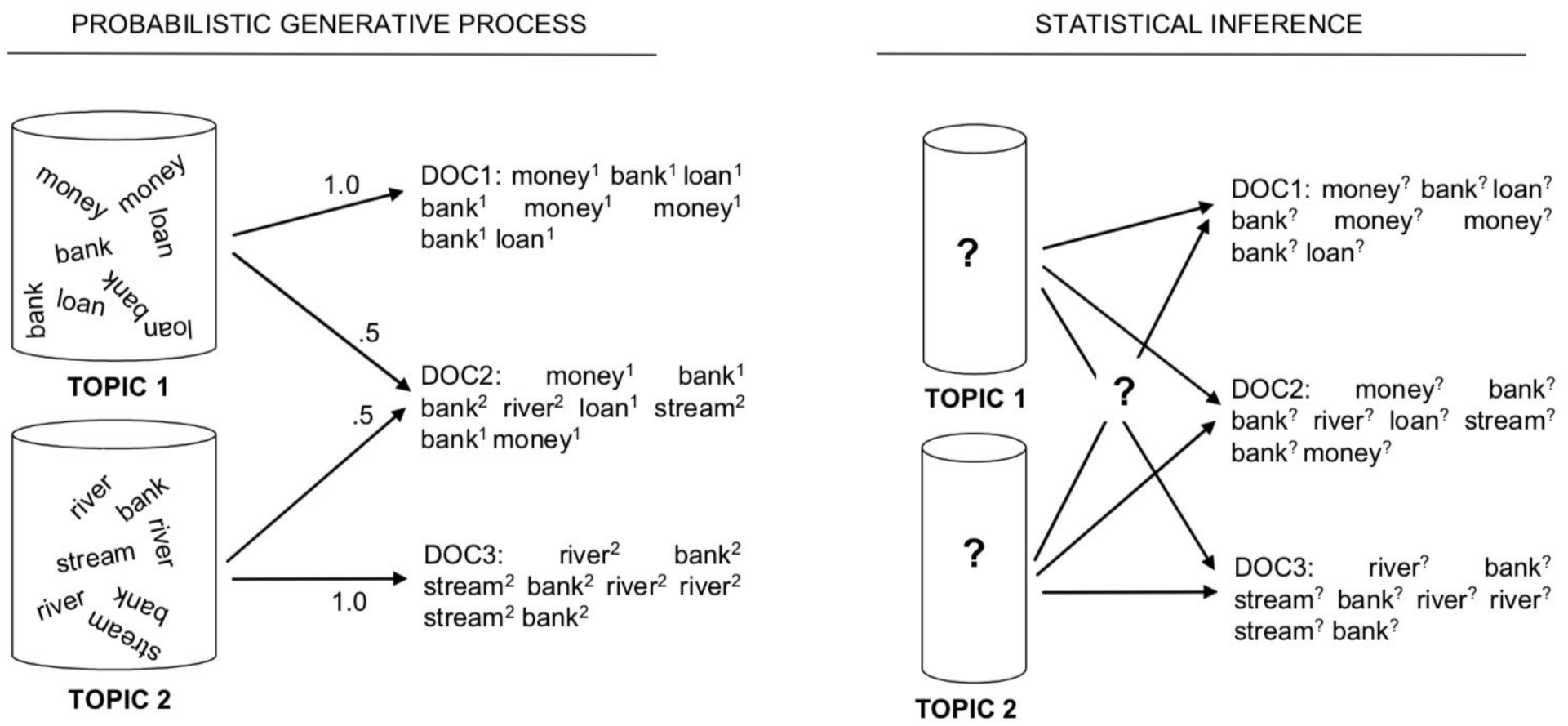
Topic modeling

- Models of a collection of composites
- Composites are documents
- Parts are words (or phrases, n-grams)
- Two outputs:
  - chance of selecting a particular part when sampling a particular topic
  - chance of selecting a particular topic when sampling a particular document or composite

Assumptions

- semantic information can be derived from a word-document co-occurrence matrix;
- topic is a probability distribution over words
- to make a new document, one chooses a distribution over topics
- for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic.
- Resulting document is a mixture of topics

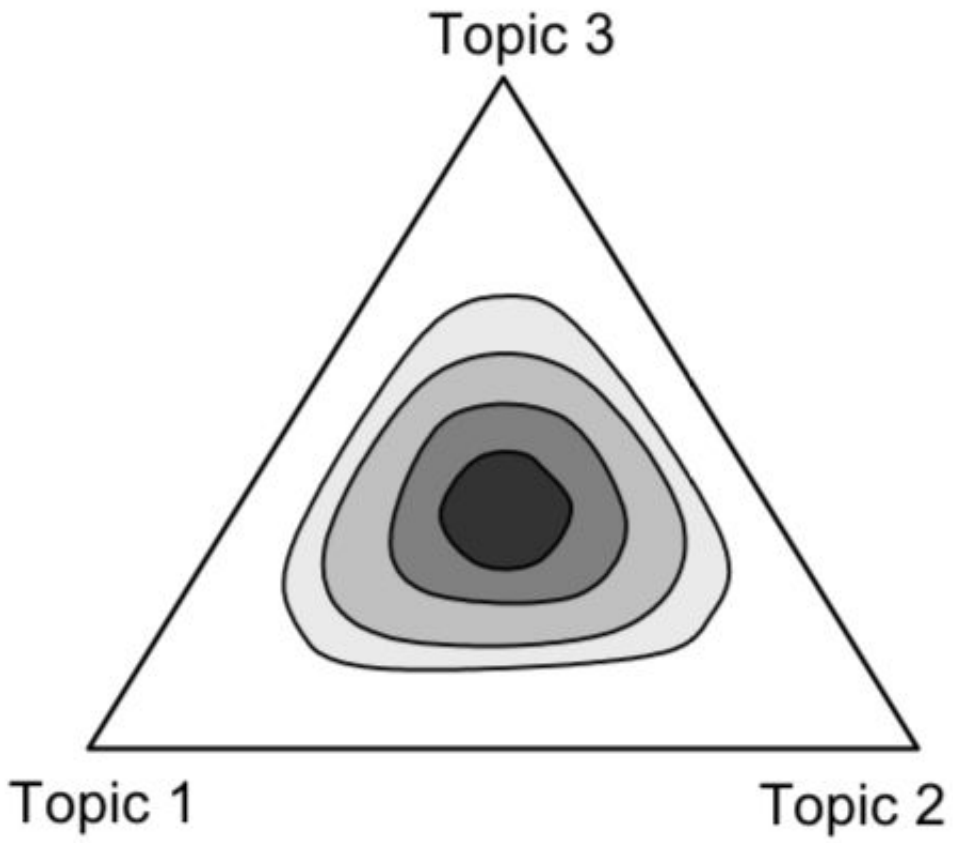
Generative model

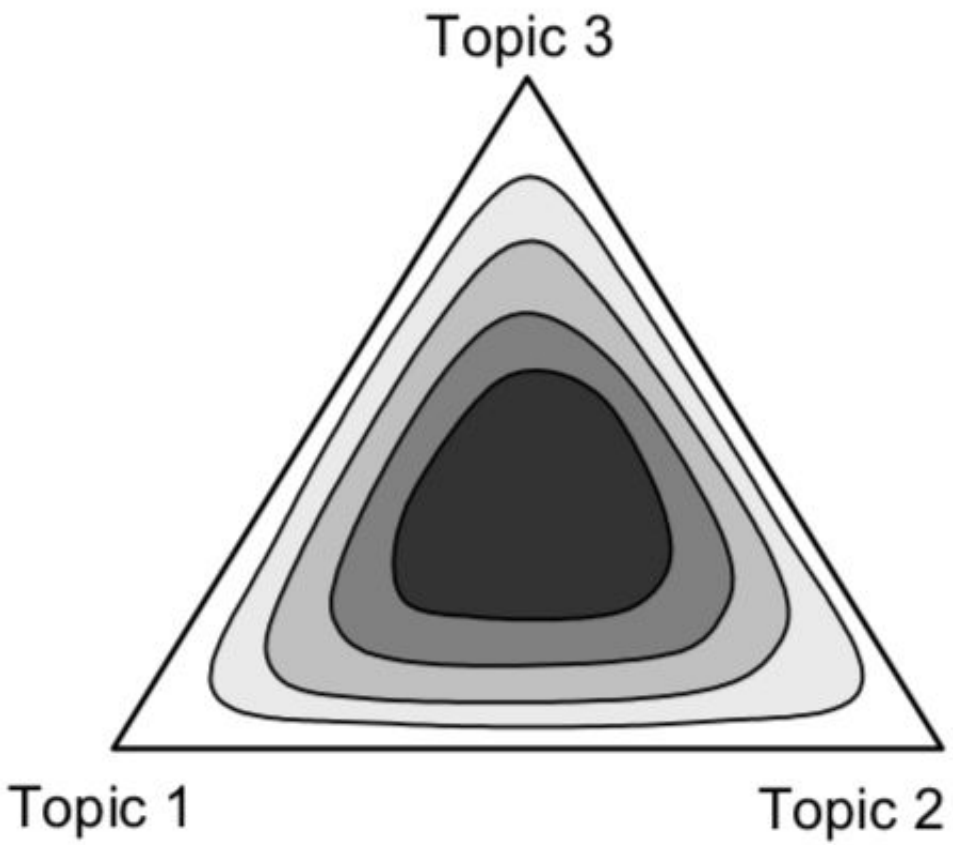


Probabilistic model

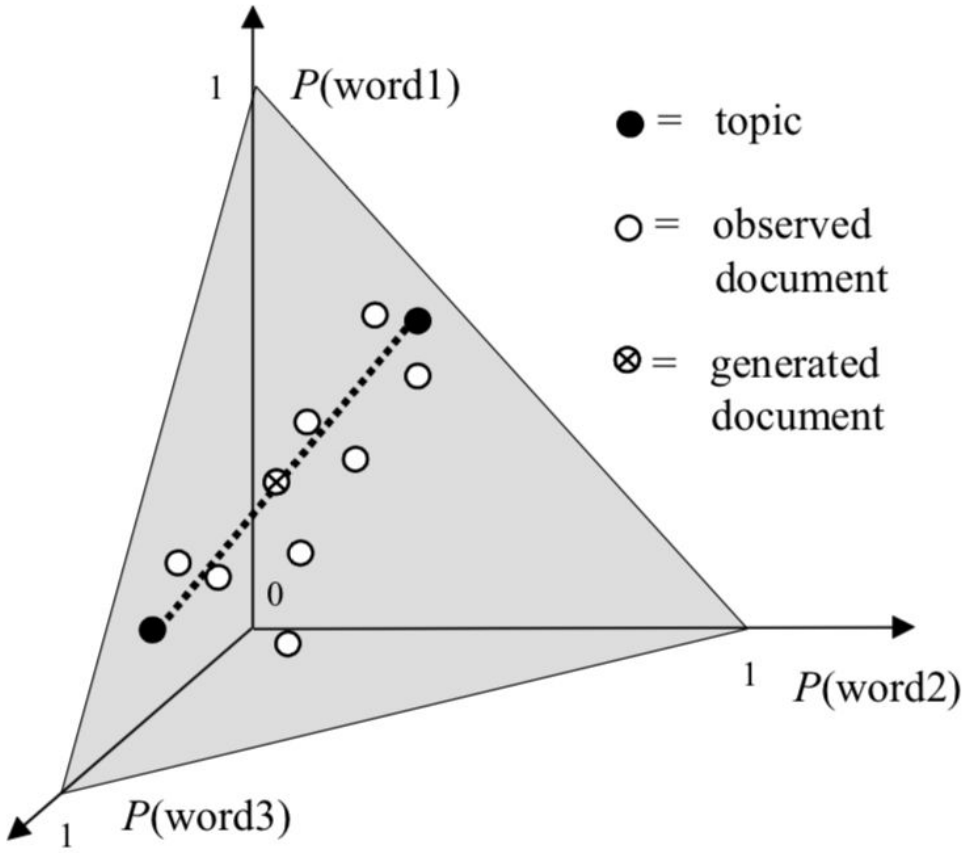
- $P(z)$ – distribution over topics *z* in a particular document
- $P(w|z)$ - the probability distribution over words *w* given topic *z*.
- first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution.
- $P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$
- T – number of topics
- Latent Dirichlet Allocation assumes a particular form of $P(z)$

Dirichlet distribution





Geometric interpretation



Main goal of the algorithm

- To invert the generative process, inferring the set of topics that were responsible for generating a collection of documents.