Topic modeling

Andrea Esuli

Topic modeling

The Latent Dirichlet Allocation (LDA) is an **unsupervised** processing tool that fits a **probabilistic generative model of text.**

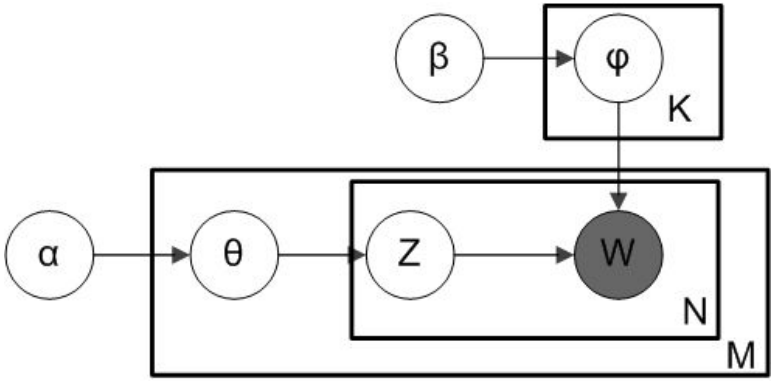
It extends a simple probabilistic language model by assuming the existence of a number of **latent (unobservable) topics.**

Every topic defines a probability law over words, i.e., a topic is a **generator of words**, following some probability distribution.

A documents is defined by a (linear) combination of topics.

Words of a document are thus generated by such weighted combination of topics.

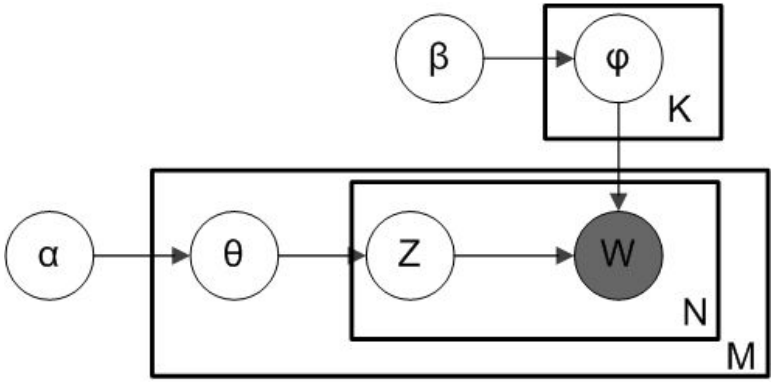
Topic modeling



The probabilistic model:

- We have a collection of M documents
- A document d_i is composed by N words w_{ij}
- α is the per-document topic distribution
- β is the per-word topic distribution
- **θ_i is the topic distribution for document i**
- **φ_k is the word distribution for topic k**
- z_{ij} is the topic of word w_j in document d_i

Topic modeling



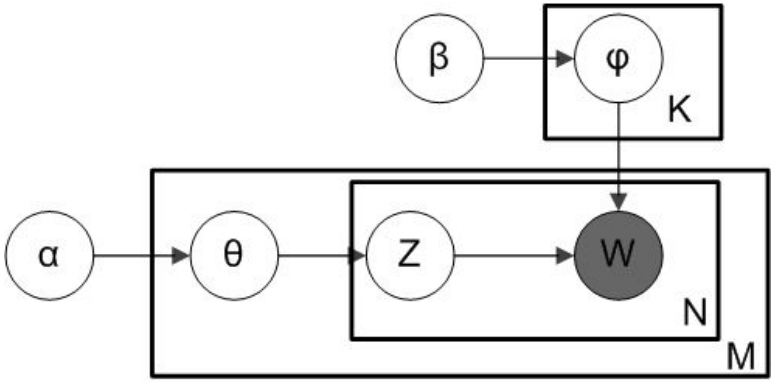
θ_i is the topic distribution for document i

Topics can be considered as **soft** clusters or a soft labeling of documents.

The vector of weights assigned to every document can be used as a **compact representation** of its content.

The argmax value can be used to produce a **hard** clustering/classification.

Topic modeling



φ_k is the word distribution for topic k

The highest values in the distribution **φ_k** for each topic defines its profile,
i.e., a list of words weighted by their relevance.

Inspection of profiles may give "meaning" to topics.

Topic modeling can be a first investigation tool to explore new collection of documents, so as to support the definition of a classification schema.