1) Intuition

The intuition behind LDA is that documents exhibit multiple topics.

LDA is a **probabilistic model of document collections** that tries to capture this intuition.

2) Generative Process

It is **most easily described by its generative process** (the way we assume the documents were generated).

- define a topic to be a distribution over a fixed vocabulary

- now for each word in the corpus, we generate the words in a two-stage process

**-** all the documents in the corpus share the **same set of topics**

**-** each document exhibits the **topics in different proportions**

**- each word** in each document is **drawn from one of the topics**

Take a look at the **graphical model**

The fact that we **only observe words and that the imagine all the rest** is reflected in the graphical model: only the W are shaded.

3) Goal

**Our goal is get back all the color in the first figure**:

- the topics

- for each document, the distribution over topics

- for each word in each document, the topic assignment

As our collective knowledge continues to be digitized and stored – in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, social networks – it becomes more difficult to find and discover that we are looking for.

We need new computational tools to help organize, search, and understand these vast amounts of information.

Right now, we work with online information using two main tools: search and links.

Imagine how much easier it would be to search and explore documents based on the themes that run through them.

That is what topic modelling is about.