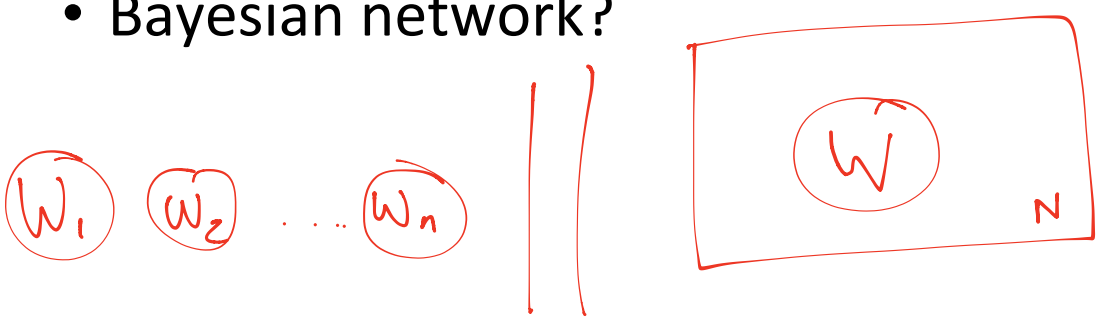
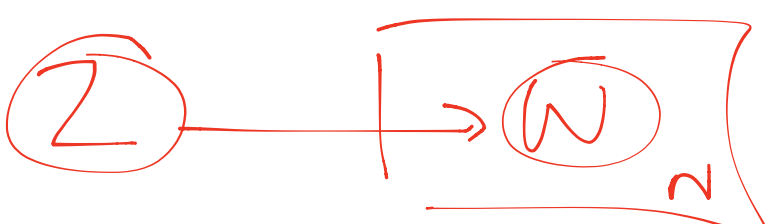
A model of text documents

- Assume for the moment that a document is made up of n words that are independently drawn from a vocabulary
- Bayesian network?



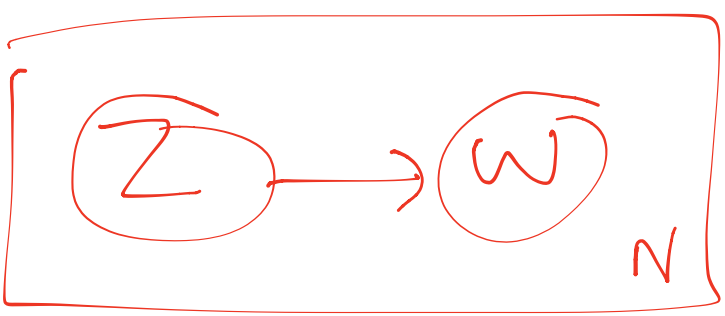
Introducing a "topic"

- Different topics will have different distributions of words
- Assume that we build a document by selecting a topic and then selecting n words.
- Bayesian model?



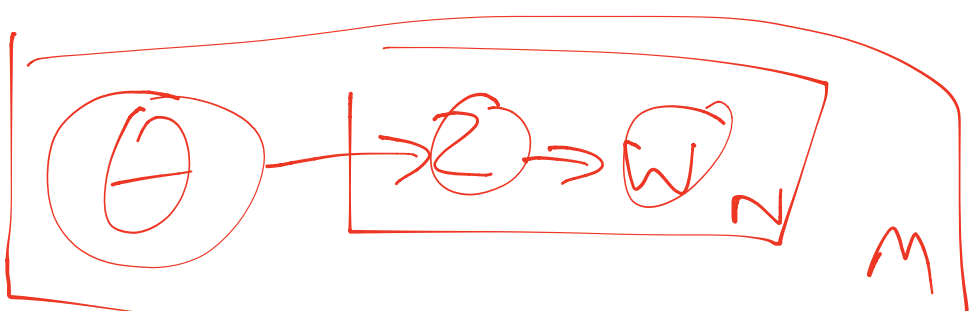
Multi-topic documents

- Let's assume now that we generate for every word position in a document first a topic, and then generate a word.
- Bayesian model?



Distributions of Topics

- How do we figure out which topic to generate from?
- This is equivalent to a probability over a multinomial probability distribution.
- The appropriate distribution is a Dirichlet distribution

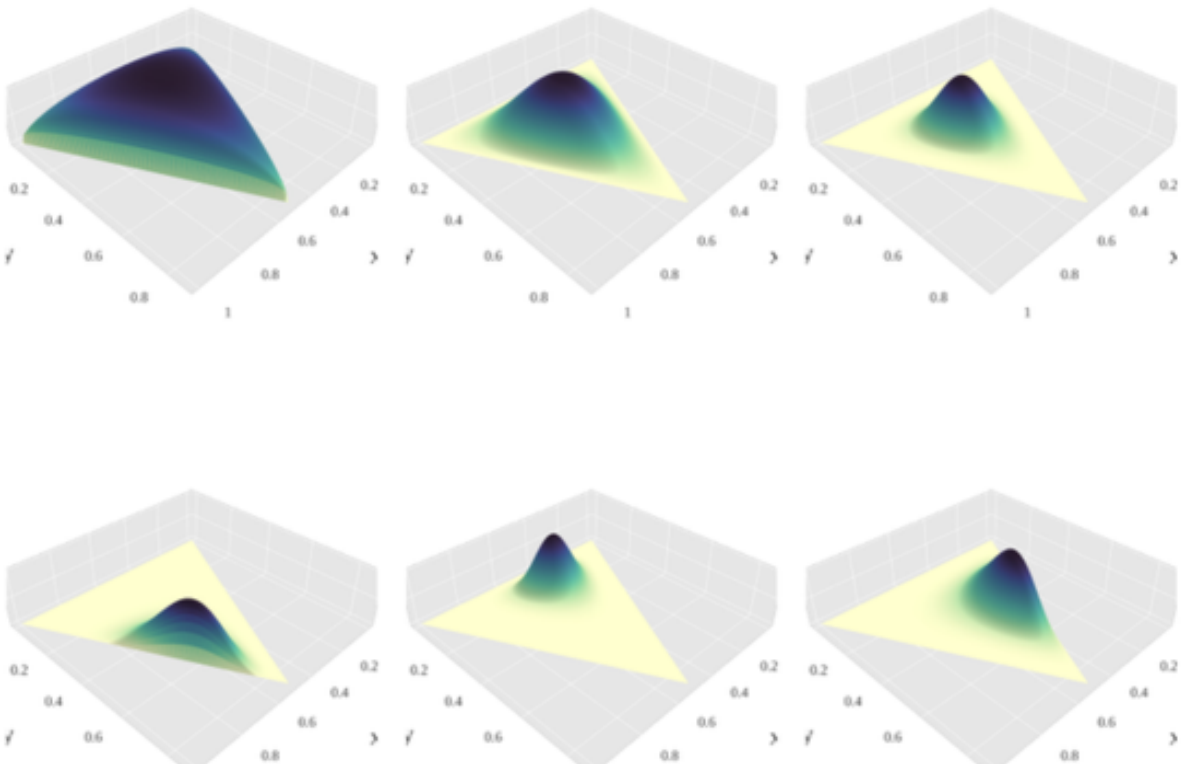


Dirichlet Distribution

$$Dirichlet(\alpha_1,...\alpha_n) = \frac{1}{B(\alpha_1,...\alpha_n)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

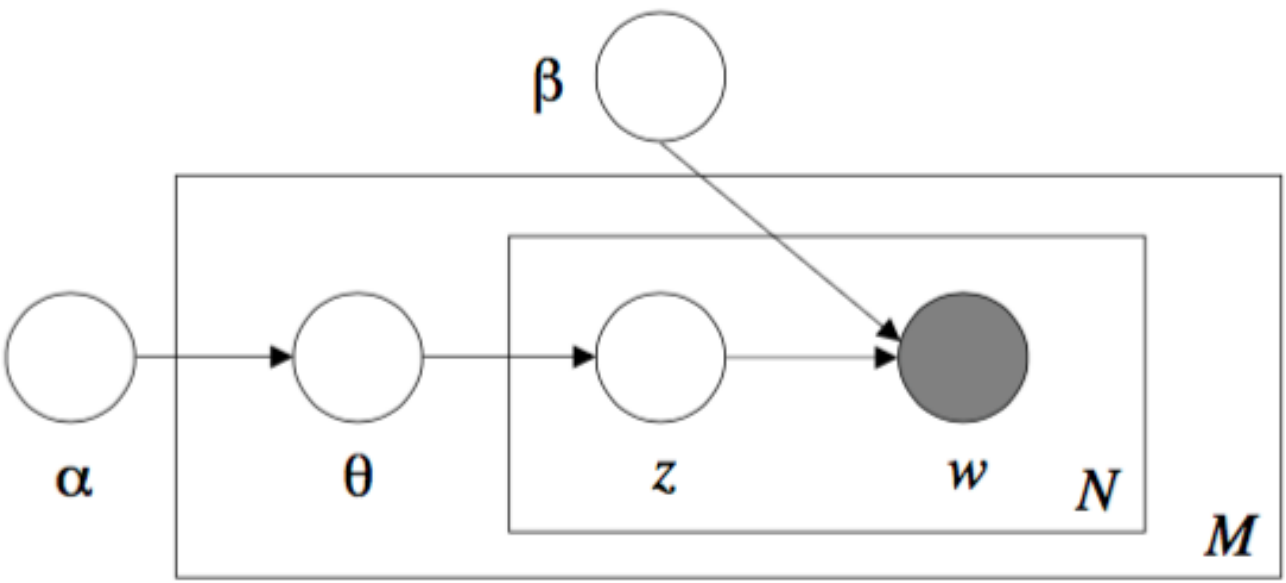
$$B(\alpha_1,...\alpha_n) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)}$$

Dirichlet for 3 dimensions



Source: Wikipedia

Final model: Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Blei, Ng, Jordan 2003