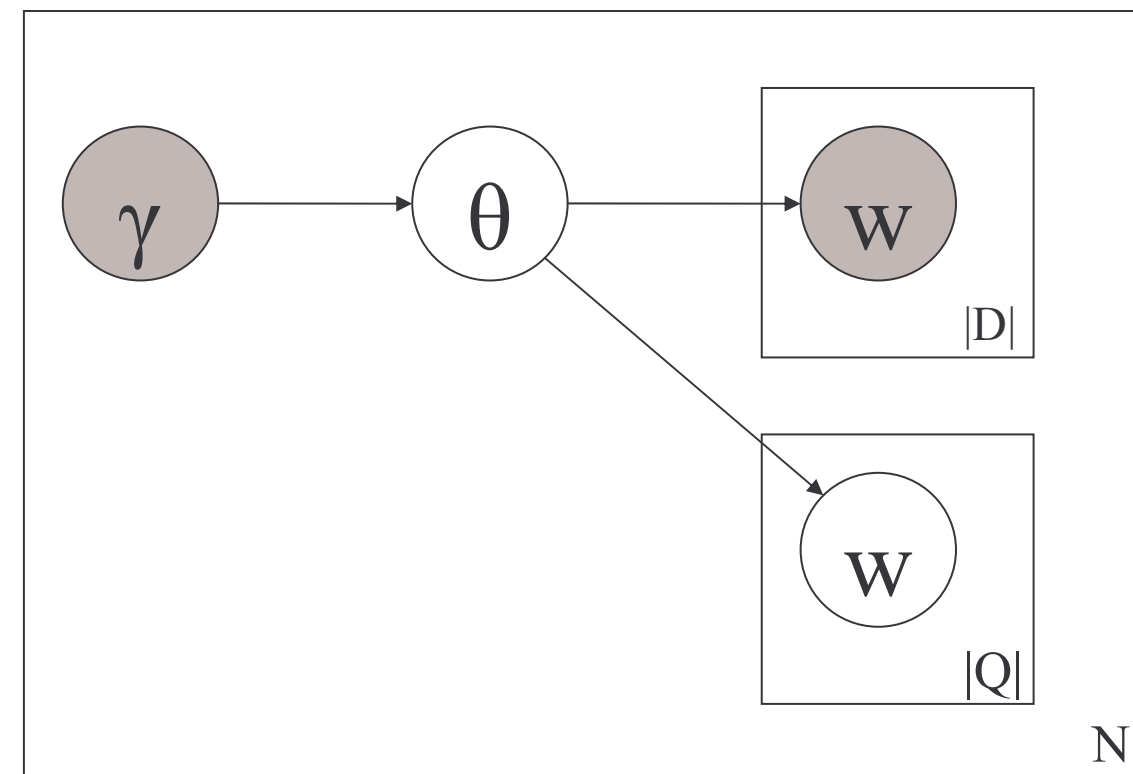


Formal Multiple-Bernoulli Models for Language Modeling

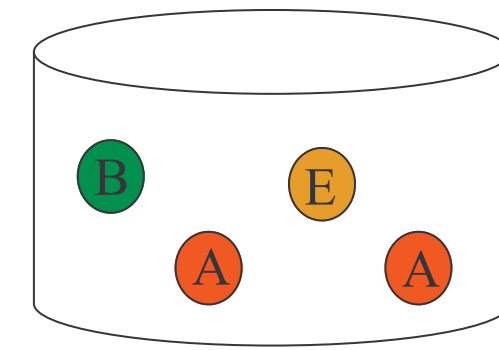
Donald Metzler, Victor Lavrenko, W.B. Croft

Multinomial



$\theta \sim \text{Dirichlet}(\gamma), w \sim \text{Multinomial}(\theta)$

- Most popular modeling assumption used in language modeling for IR
- Documents / queries represented as sequence of words
- Term presence is explicitly represented, but absence is not
- Unigram distribution over vocabulary is estimated for each document
- Documents ranked by likelihood of generating query



- term A present 2 times
- term B present 1 time
- term E present 1 time

	μ	$AvgP$
DOE	200	0.1968
WSJ	1500	0.2592

$$\gamma_w = \mu \frac{cf_w}{|C|}$$

$$P(w | D) = \frac{tf_{w,D} + \gamma_w}{|D| + \sum_w \gamma_w}$$

$$P(Q | D) = \prod_{w \in Q} P(w | D)^{q_f}$$

Why Multiple-Bernoulli?

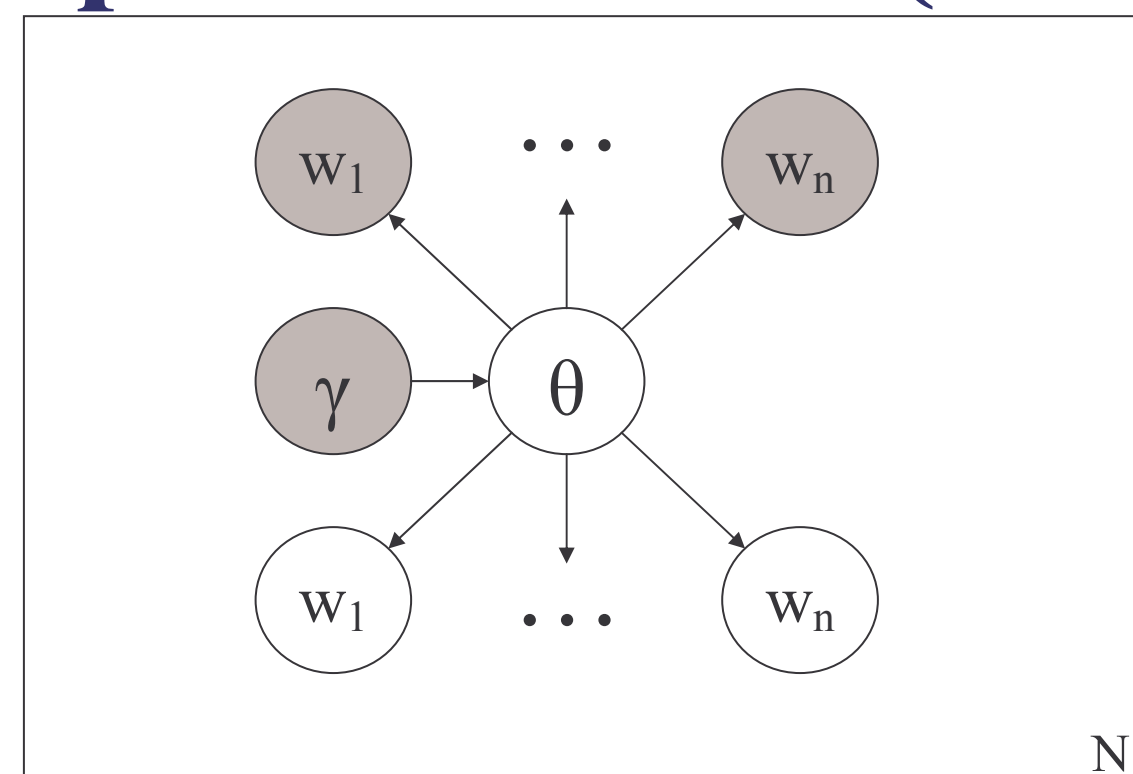
“NOT”

- Multinomial model only models presence of terms
- No intuitive interpretation of how likely it is a term is absent from a document
- Under the Multiple-Bernoulli model, the probability a term w is absent is simply $1 - P(w | D)$

Arbitrary Binary Features

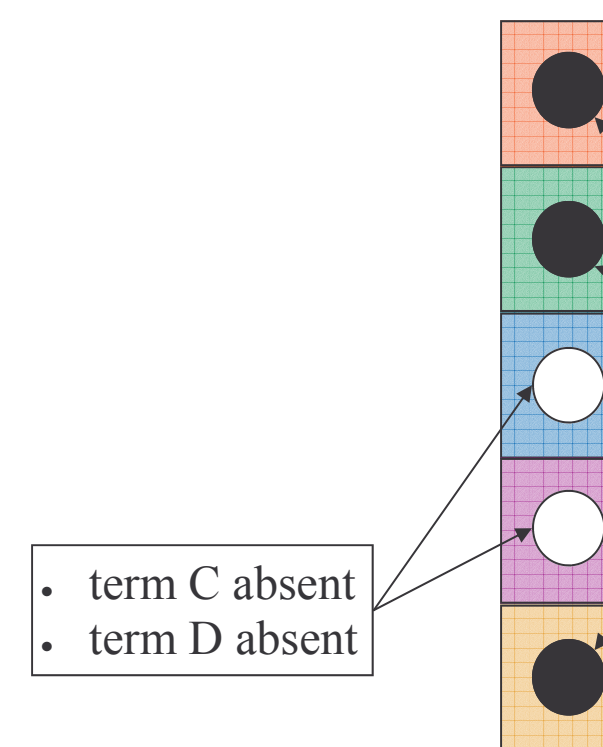
- Multinomial language models are typically distributions over single terms
- Can model arbitrary features (phrases, term markup, etc) with multinomial, but results in ballooned event space which makes estimation more difficult
- Multiple-Bernoulli allows straightforward incorporation of arbitrary binary features that can be easily extracted from both the document and query
- Phrases now accessible in language modeling framework

Multiple-Bernoulli (Model A)



$\theta \sim \text{Multi-Beta}(\gamma_\alpha, \gamma_\beta), w_i \sim \text{Bernoulli}(\theta)$

- Based on original language modeling for IR model introduced by Ponte and Croft that used somewhat *ad hoc* estimation
- Our model uses more formal Bayesian estimation techniques
- Documents / queries represented as a single binary vector
- Term presence / absence is explicitly modeled
- Multiple term occurrences cannot be represented



- term A present
- term B present
- term E present

- term C absent
- term D absent

	μ	$AvgP$
DOE	10	0.1616
WSJ	500	0.1050

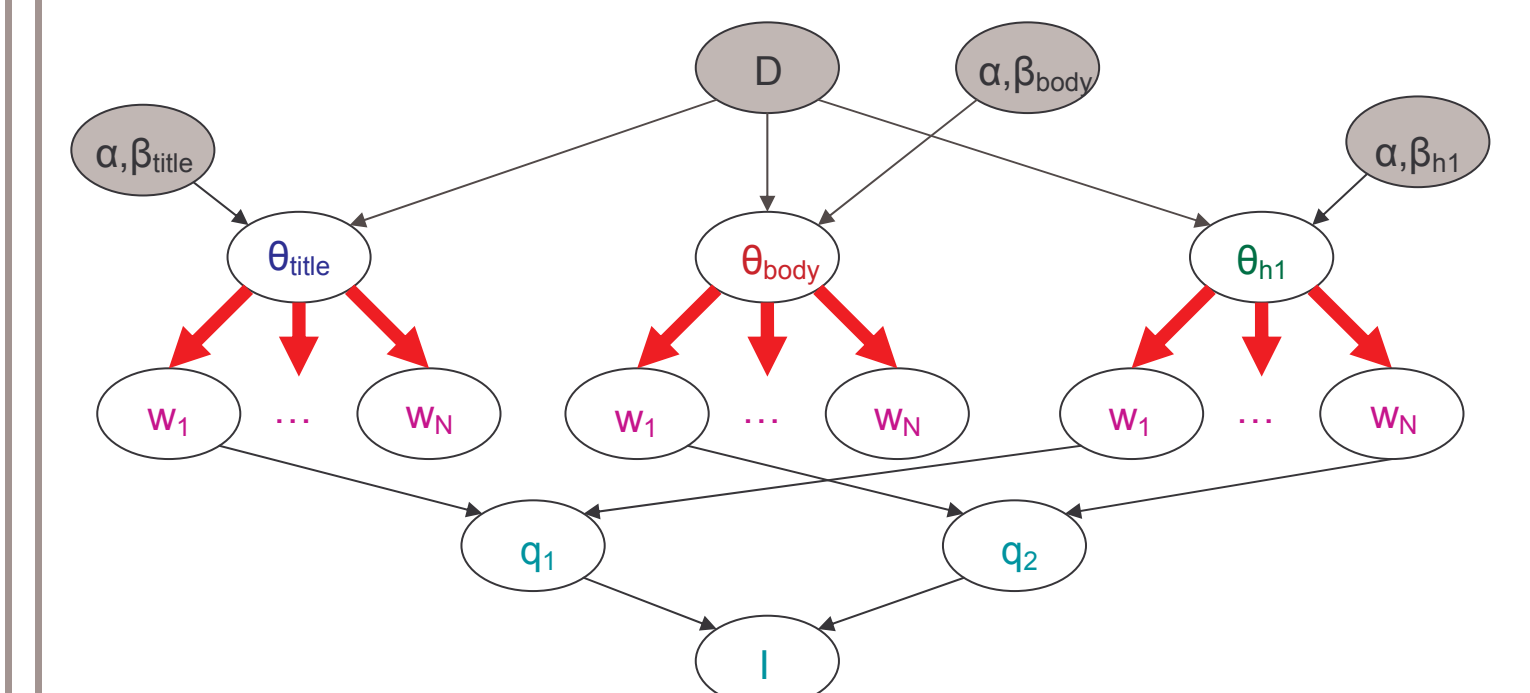
$$\gamma_{\alpha,w} = \mu \frac{df_w}{N}, \gamma_{\beta,w} = \mu \left(1 - \frac{df_w}{N}\right)$$

$$P(w | D) = \frac{df_w + \gamma_{\alpha,w}}{1 + \gamma_{\alpha,w} + \gamma_{\beta,w}}$$

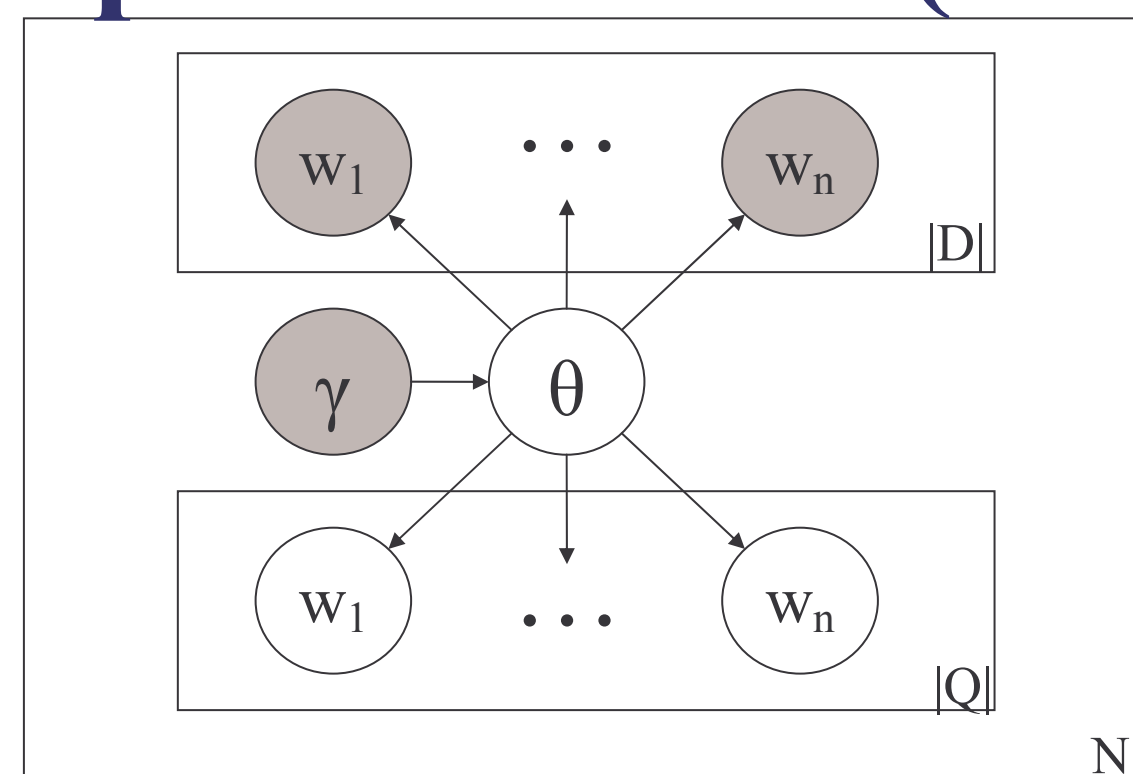
$$P(Q | D) = \prod_{w \in Q} P(w | D) \prod_{w \notin Q} (1 - P(w | D))$$

Inference Network Framework Integration

- Can combine inference network and language modeling approaches to retrieval into a single model
- Use multiple-Bernoulli language model estimates in place of *tf.idf* estimates for probabilities within network
- Currently developing the INDRI search engine based on this framework



Multiple-Bernoulli (Model B)

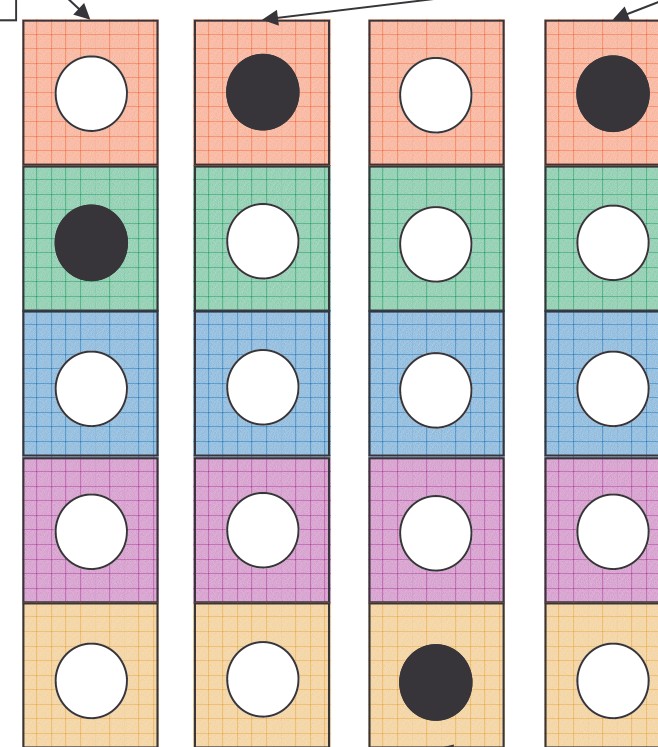


$\theta \sim \text{Multi-Beta}(\gamma_\alpha, \gamma_\beta), w_i \sim \text{Bernoulli}(\theta)$

- Similar to Model A, but makes use of richer document / query representation
- Documents are represented as set of binary feature vectors
- Simplest case: indicator vectors for each term in document
- More complex: arbitrary binary features extracted from document
- Akin to a naive-Bayes classifier

term B feature extracted once

term A feature extracted twice



term E feature extracted once

	μ	$AvgP$
DOE	100	0.1966
WSJ	2000	0.2540

$$\gamma_{\alpha,w} = \mu \frac{cf_w}{|C|}, \gamma_{\beta,w} = \mu \left(1 - \frac{cf_w}{|C|}\right)$$

$$P(w | D) = \frac{tf_{w,D} + \gamma_{\alpha,w}}{|D| + \gamma_{\alpha,w} + \gamma_{\beta,w}}$$

$$P(Q | D) = \prod_{w \in Q} P(w | D)^{q_f} \prod_{w \notin Q} (1 - P(w | D))^{q_f - q_w}$$