Help System

- The domain of H is the set of all help pages.
  The observations are the words in the query.
- What probabilities are needed?
  What pseudo-counts and counts are used?
  What data can be used to learn from?

Help System

- Suppose the help pages are $\{h_1, \ldots, h_k\}$.
- Words are $\{w_1, \ldots, w_m\}$.
- Bayes net requires:
  - $P(h_i)$, these sum to 1 ($\sum_i P(h_i) = 1$)
  - $P(w_j \mid h_i)$, these do not sum to one
- Maintain "counts" (pseudo counts + observed cases):
  - c_i the number of times h_i was the correct help page
  - $s = \sum_i c_i$
  - u_{ij} the number of times h_i was the correct help page and word w_j was used in the query.
- $P(h_i) = c_i/s$
- $P(w_j \mid h_i) = u_{ij}/c_i$

Learning + Inference

- Q is the set of words in the query.
- Learning: if h_i is the correct page: Increment s, c_i and u_{ij} for each $w_j \in Q$.
- Inference:

$$P(h_i \mid Q) \propto P(h_i) \prod_{w_j \in Q} P(w_j \mid h_i) \prod_{w_j \notin Q} (1 - P(w_j \mid h_i))$$

$$\left\langle \begin{matrix} \textit{expensive} \\ \textit{inference} \end{matrix} \right\rangle = \frac{c_i}{s} \prod_{w_j \in Q} \frac{u_{ij}}{c_i} \prod_{w_j \notin Q} \frac{c_i - u_{ij}}{c_i}$$

Learning + Inference

- Q is the set of words in the query.
- Learning: if h_i is the correct page: Increment s, c_i and u_{ij} for each $w_j \in Q$.
- Inference:

$$P(h_i \mid Q) \propto P(h_i) \prod_{w_j \in Q} P(w_j \mid h_i) \prod_{w_j \notin Q} (1 - P(w_j \mid h_i))$$

$$\left\langle \begin{array}{c} \textit{expensive} \\ \textit{inference} \end{array} \right\rangle = \frac{c_i}{s} \prod_{w_j \in Q} \frac{u_{ij}}{c_i} \prod_{w_j \notin Q} \frac{c_i - u_{ij}}{c_i}$$

$$= \frac{c_i}{s} \prod_{w_j} \frac{c_i - u_{ij}}{c_i} \prod_{w_j \in Q} \frac{u_{ij}}{c_i - u_{ij}}$$

$$\left\langle \begin{array}{c} \textit{expensive} \\ \textit{learning} \end{array} \right\rangle = \Psi_i \prod_{w_j \in Q} \frac{u_{ij}}{c_i - u_{ij}}$$

Issues

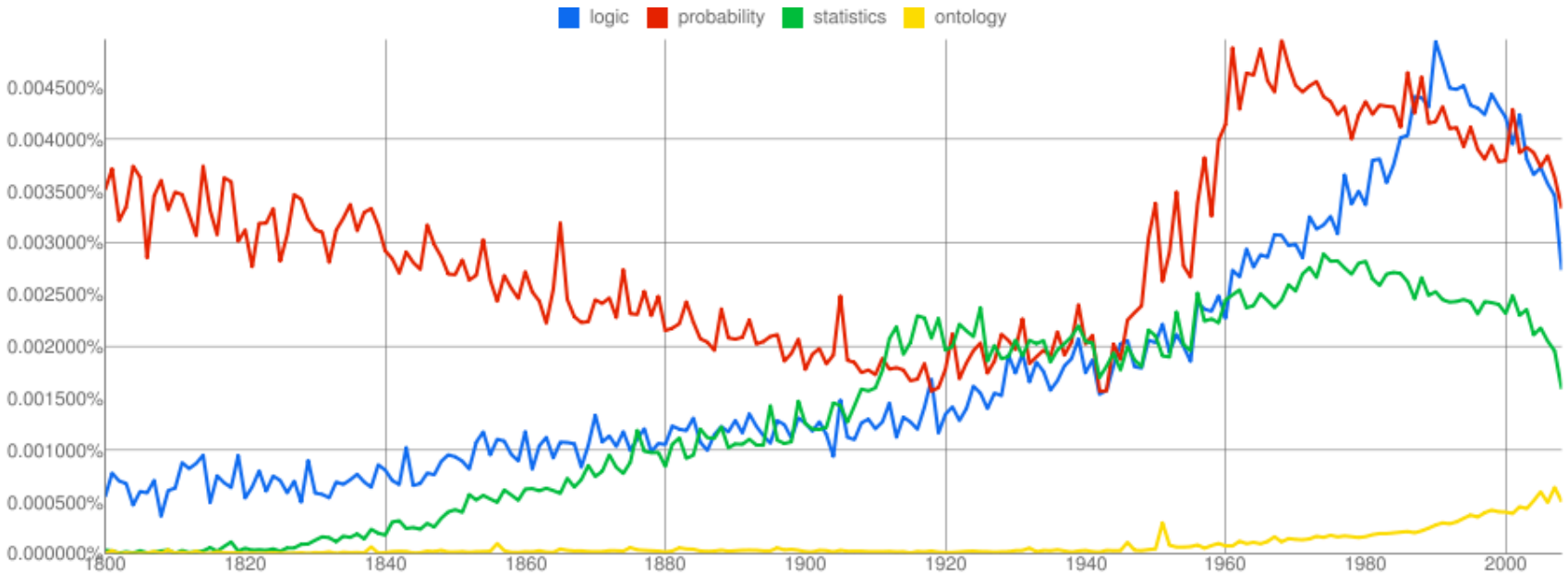
- What if the most likely page isn't the correct page?
- What if the user can't find the correct page?
- What if the user mistakenly thinks they have the correct page?
- Can some pages never be found?
- What about common words?
- What about words that affect the meaning, e.g. "not"?
- What about new words?
- What do we do with new help pages?
- How can we transfer the language model to a new help system?

Simple Language Models

Sentence is a sequence of words: $w_1, w_1, w_3, \ldots$.
Modeled as:

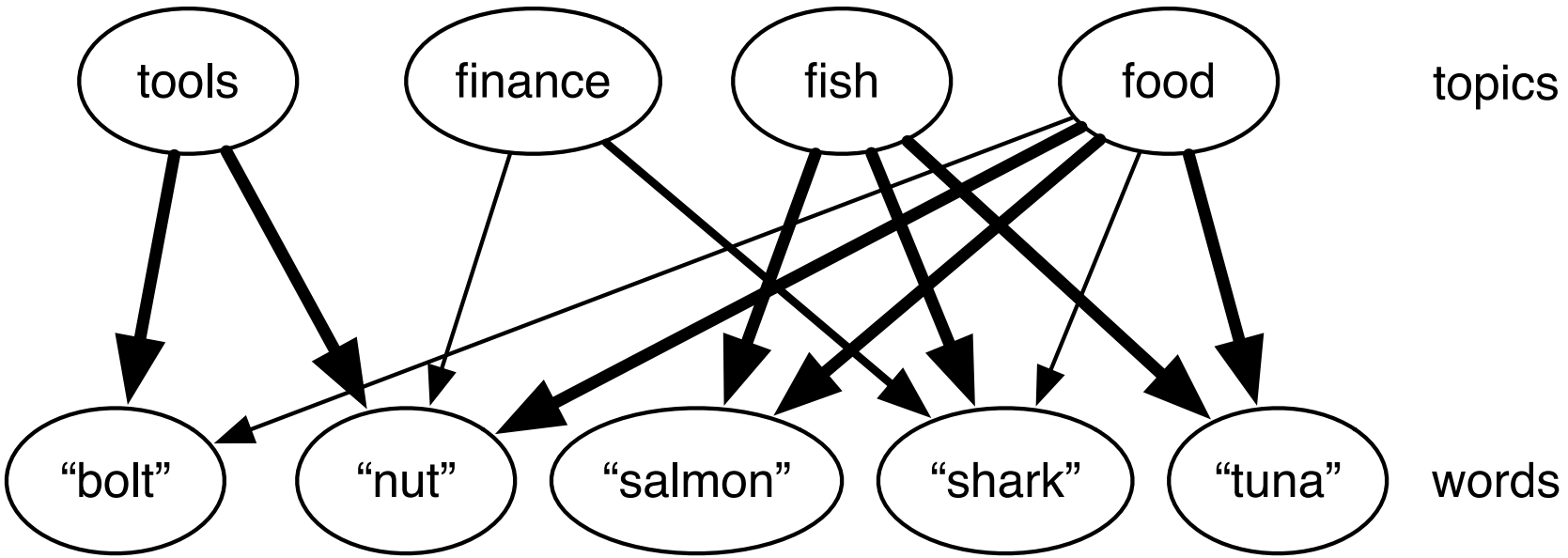
- Set of words
- Unigram (bag of words): Model $P(W_i)$.
  Range of W_i is the set of all words.
- Bigram: Model $P(W_i \mid W_{i-1})$
- Trigram: Model $P(W_i \mid W_{i-1}, W_{i-2})$

Logic, Probability, Statistics, Ontology over time

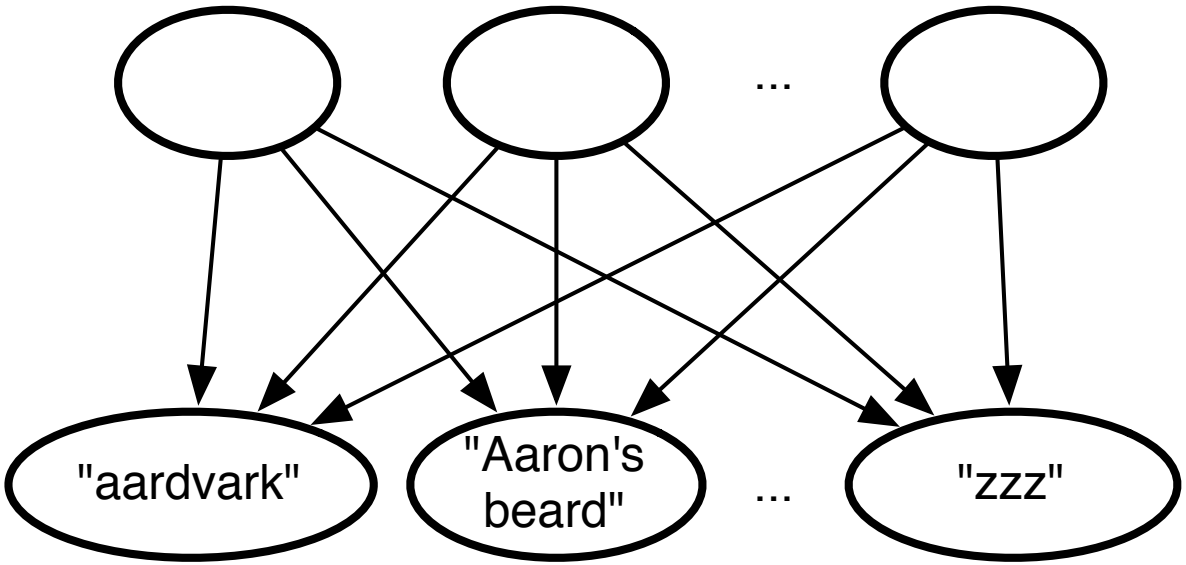


From: Google Books Ngram Viewer
(http://books.google.com/ngrams/)

Topic Model


tools
finance
fish
food
topics
“bolt”
“nut”
“salmon”
“shark”
“tuna”
words

Google's rephil



900,000 topics

350,000,000 links

12,000,000 words