Modeling Document Collections

David Mareček

October 19, 2022




Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)

Modelling text documents

GOP abortion foes are criminalizing the doctor-patient relationship
"The doctor-patient relationship." For more than 20 years, conservative propagandists and their Republican allies have used that four-word bludgeon to beat back universal health care reform. In 1994, GOP strategist Bill Kristol warned that "the Clinton Plan is damaging to the quality of American medicine and to the relationship between the patient and the doctor." Kristol's successful crusade to derail Bill Clinton's reform effort was greatly aided by future "death panels" fabulist Betsy McCaughey, who wrongly warned that Americans would even lose the right to see the doctor of their choice. Twelve years later, President George W. Bush proclaimed, "Ours is a party that understands the best health care system is when the doctor-patient relationship is central to decision-making."

How would we model this document?

- Unigram model (bag of words): $p = \prod_{i=1}^{N} p(w_i)$
- Bigram model: $p = \prod_{i=1}^{n} p(w_i|w_{i-1})$
- ...

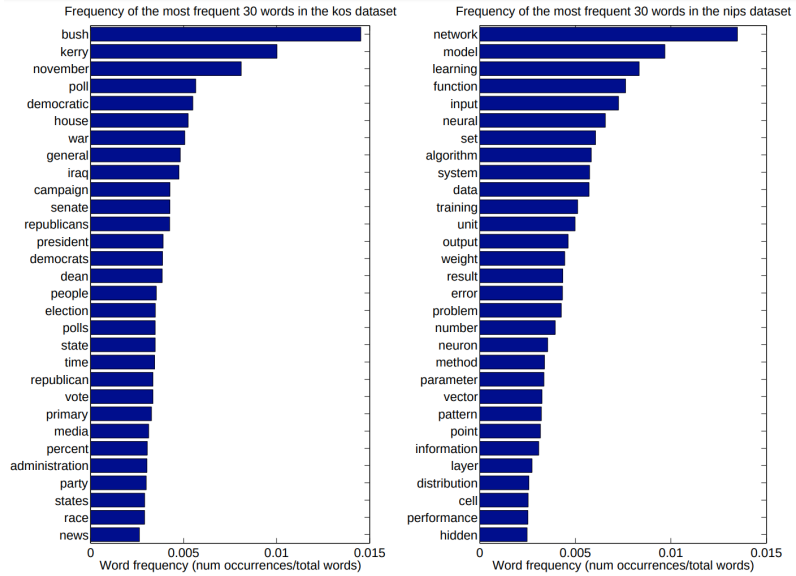
Example: word counts in text

Consider describing a text document by the frequency of occurrence of every distinct word (bag of words model).

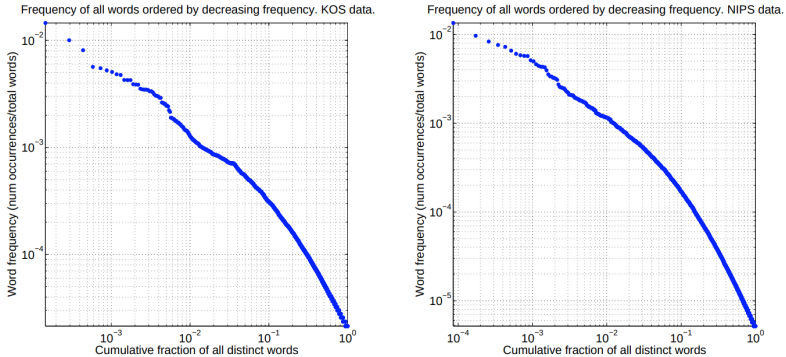
For illustration consider two collections of documents:

- KOS (political blog — http://dailykos.com):
  - $D = 3,430$ documents (blog posts)
  - $n = 353,160$ words
  - $m = 6,906$ distinct words
- NIPS (machine learning conference — http://nips.cc):
  - $D = 1,500$ documents (conference papers)
  - $n = 746,316$ words
  - $m = 12,375$ distinct words

Example: word counts in text



Zipf's law: different collections, similar behavior



Zipf's law states that the frequency of any word is inversely proportional to its rank in the frequency table.

$$\exists k: \quad frequency \approx \frac{k}{rank} \quad \Longrightarrow \quad log(frequency) \approx log(k) - log(rank)$$

Automatic Categorisation of Documents

Can we make use of the statistical distribution of words, to build an automatic document categorisation system?

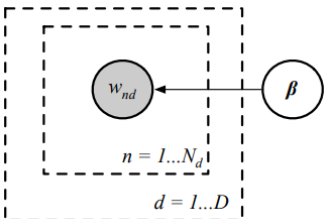
- The learning system would have to be unsupervised
- We don't a priori know what categories of documents exist
- It must automatically discover the structure of the document collection.
- What should it even mean, that a document belongs to a category, or has certain properties?

How can we design such a system?

A really simple document model

Consider a collection of D documents from a vocabulary of M words.

- N_d: number of words in document d.
- w_{nd}: n-th word in document d ($w_{nd} \in 1 \dots M$).
- $w_{nd} \sim Cat(\beta)$: each word is drawn from a discrete categorical distribution with parameters β
- $\beta = [\beta_1, \dots, \beta_M]$ parameters of a categorical/multinomial distribution over the M vocabulary words.



A really simple document model

We can fit β by maximizing the likelihood:

$$\beta = \arg\max_{\beta} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \text{Cat}(w_{nd}|\beta) = \arg\max_{\beta} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{w_{nd}}$$

$$\beta = \arg\max_{\beta} \text{Mult}(c_1, \ldots, c_M|\beta, N) = \arg\max_{\beta} \prod_{m=1}^{M} \beta_m^{c_m}$$

$$\beta_m = \frac{c_m}{N} = \frac{c_m}{\sum_{l=1}^{M} c_l}$$

- $N = \sum_{d=1}^{D} N_d$: total number of words in the collection.
- $c_m = \sum_{d=1}^{D} \sum_{n=1}^{N_d} I(w_{nd} = m)$: total count of vocabulary word m.

Lagrange Multipliers

We want to find the maximum or minimum of a function $f(x)$ subjected to some equality constraint $g(x) = 0$.

We form the Lagrangian function $L(x, \lambda) = f(x) - \lambda g(x)$.

The solution corresponding to the original constrained optimization is always a saddle point of the Lagrangian function.

In our case, the equality constraint is:

$$g(\beta) = \sum_{m=1}^{M} \beta_m - 1 = 0$$

We want to maximize the (log) likelihood (probability of data)

$$\log \prod_{m=1}^{M} \beta_m^{c_m} = \sum_{m=1}^{M} c_m \log \beta_m$$

Maximum Likelihood for Multinomial Distribution

We form the following Lagrangian function:

$$L(\beta, \lambda) = \sum_{m=1}^{M} c_m \log \beta_m + \lambda(1 - \sum_{m=1}^{M} \beta_m)$$

We take derivatives of the Lagrangian function.

By setting them to zero, we obtain

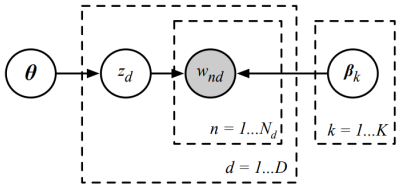
$$\frac{\partial L}{\partial \beta_m} = \frac{c_m}{\beta_m} - \lambda = 0 \quad \Rightarrow \quad \beta_m = \frac{c_m}{\lambda}$$

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{m=1}^{M} \beta_m = 0 \quad \Rightarrow \quad \sum_{m=1}^{M} \frac{c_m}{\lambda} = \frac{n}{\lambda} = 1 \quad \Rightarrow \quad n = \lambda \quad \Rightarrow \quad \beta_m = \frac{c_m}{n}$$

Limitations of the really simple document model

- Document d is the result of sampling N_d words from the categorical distribution with parameters β.
- β estimated by maximum likelihood reflects the aggregation of all documents.
- All documents are therefore modelled by the global word frequency distribution.
- This generative model does not specialise.
- We would like a model where different documents might be about different topics.

A mixture of categoricals model



$$z_d \sim Cat(\theta)$$
$$w_{nd}|z_d \sim Cat(\beta_{z_d})$$

We want to allow for a mixture of K categoricals parametrised by $\beta_1, \ldots, \beta_K$.
Each of those categorical distributions corresponds to a document category.

- $z_d \in 1, \ldots, K$ assigns document d to one of the K categories.
- $\theta_k = p(z_d = k)$ is the probability any document d is assigned to category k.
- so $\theta = [\theta_1, \ldots, \theta_K]$ is the parameter of a categorical distribution over K categories.

We have introduced a new set of hidden variables z_d.

- How do we fit those variables? What do we do with them?
- Are these variables interesting? Or are we only interested in θ and β?