Automatic Labeling for Multinomial Topic Models

Problem: meaningful labels for discovered topics. w/o being subjective through manual labeling *

Solution: Optimization problem involving minimizing Kullback-Leiber divergence while maximizing mutual information between a label and a topic model.

Href: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.923.5020&rep=rep1&type=pdf

→ Try to minimize distance b/w label "ℓ" and topic "θ"

Kullback-Leiber Definition:

Divergence ⇝ $$D_{KL}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \neq D_{KL}(Q \| P)$$

$x \in \mathcal{X}$ → Probability Space

$\{\text{trump, ukraine, zelensky, putin}\} = \theta$

$L = \begin{Bmatrix} \ell_1 = \text{"war on ukraine"} \\ \ell_2 = \text{"sports game"} \\ \ell_3 = \text{"cooking show"} \end{Bmatrix}$

then $D_{KL}(\theta \| \ell_1) > D_{KL}(\theta \| \ell_2)$

Href: https://en.wikipedia.org/wiki/Kullback–Leibler_divergence

Context: topic model θ in text collection C is a probability distribution of words ω in vocab set V:

$$\sum_{\omega \in V} P(\omega | \theta) = 1$$

topic label : sequence of words which is semantically meaningful and covers the latent meaning of topic θ.

Relevance Score : $s(l, \theta)$ measures semantic similarity between the label and the topic model.

relative to long or logic

To generate labels that are understandable, semantically relevant, discriminative across topics, and of high coverage of each topic, we first extract a set of candidate labels in a preprocessing step.*

↳ Best labels need to be phrases present within the collection C

"Candidate label"

Phrase Generation:

→ Chunking / Shallow Parsing ~ Aims to identify short phrases in text using parts of speech tags to make decisions of chunking according to some grammar.

- look for chunks / phrases frequently appearing in text as candidate labels

↳ Ngram Testing: extract meaningful phrases from word ngrams based on Statistical tests.

↳ if words in an ngram co-occur w/ eachother, the n-gram is more likely to be an n-word phrase.

Href: https://medium.com/analytics-vidhya/generating-meaningful-phrases-from-unstructured-news-data-d4e217a7da43

Semantic Relevance Scoring

↳ The zero-order Relevance

any reasonable measure of relevance b/w θ and ℓ should compare ℓ to the distribution of words that define θ.

θ is a topic

$$\text{Score} = \log \frac{P(\ell \mid \theta)}{P(\ell)} = \sum_{0 \le i \le m} \log \frac{P(u_i \mid \theta)}{P(u_i)}$$

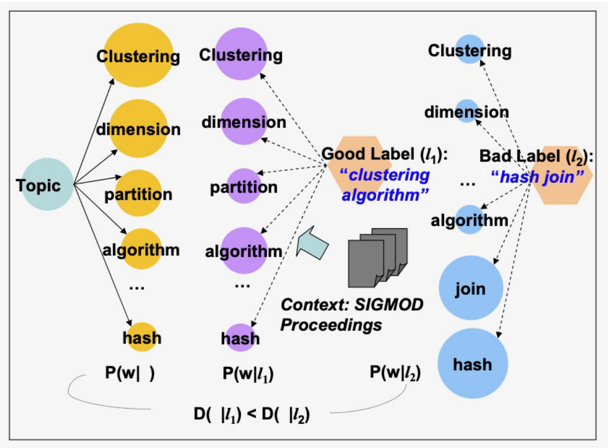
↳ u_i is a word

ℓ is a phrase where $\ell = u_0 u_1 \ldots u_m$

- a phrase containing more "important" (high $P(\omega \mid \theta)$) words in the topic distribution is assumed to be a good label.
- Does not look at context information from reference collection

↳ The first-order Relevance

- The semantics of a topic model should be interpreted in a context.
- A natural context to interpret a topic is the original collection from which the topic model is extracted
- Represent a candidate label also with a multinomial distribution of words. → $\{P(w|l)\}$ decided by l



can measure closeness of $\{P(w|l)\}$ and $\{P(w|\theta)\}$ using Kullback-Leiber divergence $D(\theta \| l)$

- To use this relevance score, we need to approximate distribution $\{P(w|l)\}$ ⤳ include a context collection C.
  ↳ substitute $\{P(w|l)\}$ w/ $\{P(w|l, C)\}$

↳ Relevance Scoring Function

$$Score(l,\theta) = -D(\theta \| l) = -\sum_w p(w|\theta) \log \frac{p(w|\theta)}{p(w|l)}$$

$$= -\sum_w p(w|\theta) \log \frac{p(w|C)}{p(w|l,C)} - \sum_w p(w|\theta) \log \frac{p(w|\theta)}{p(w|C)} - \sum_w p(w|\theta) \log \frac{p(w|l,C)}{p(w|l)}$$

$$= \sum_w p(w|\theta) \log \frac{p(w,l|C)}{p(w|C)\,p(l|C)} - D(\theta\|C) - \sum_w p(w|\theta) \log \frac{p(w|l,C)}{p(w|l)}$$

$$= \sum_w p(w|\theta)\, PMI(w,l|C) - D(\theta\|C) + Bias(l,C)$$

- can ignore b/ws if all candidate labels are generated from collection C.

Divergence b/w the topic and the labeling context.
- identical if all candidate labels came from same collection

↳ Bias of using Context C to infer the semantic relevance of l and θ.
- can be utilized to incorporate priors of candidate labels.

- First component can be written as the expectation of pointwise mutual information b/w l and the terms in the topic model given the context C $E_\theta(PMI(w,l|C))$.

- PMI of x,y is correlation b/w events x and y

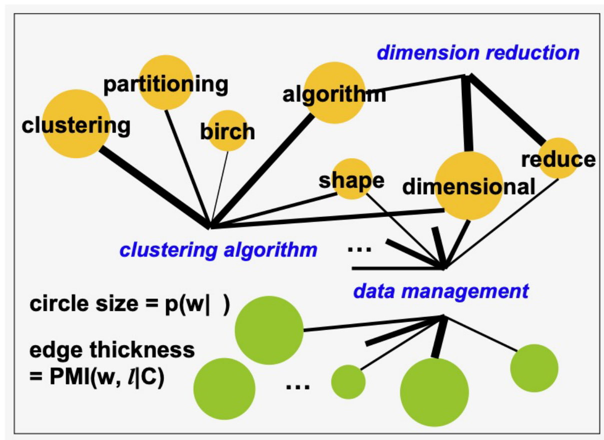
$= \log(P(x,y)/P(x)P(y))$

• PMI (w, l | C) can all be precomputed independently of the topic models to be labeled.

* use Laplace smoothing to filter out "P(w, l|C) = 0" words in order to avoid PMI (w, l|C) being undefined

↳ Intuitive interpretation

• can construct weighted graph where each node is either a term in a topic model or a candidate label.



• Each edge b/w label and topical term measured using PMI(w, l|C)

• The weight of each node indicates the importance of the term to this topic. W.e. The weight of each edge indicates how strongly the label and the term are semantically associated.

• The scoring function $E_\theta(PMI(w, l|C))$ would rank a label node higher if it generally has stronger semantic relation to those important topical words.

High Coverage Labels