Precision-Recall Balanced Topic Modelling

Seppo Virtanen[1] and Mark Girolami[1,2]

University of Cambridge[1] The Alan Turing Institute[2]

Summary

Topic models are becoming increasingly relevant probabilistic models for dimensionality reduction of text data, inferring topics that capture meaningful themes of frequently co-occurring terms.
In this work, we present:

- New insights into topic modelling from an information retrieval perspective.
- New evaluation measures for topic modelling based on the precision-recall trade-off.
- Novel topic model combined with an efficient inference algorithm that allows the user to balance between contributions of precision and recall, inferring more coherent and meaningful topics.

Topic models are recall-biased

M documents $\boldsymbol{y}_m$, where $m = 1, \ldots, M$, such that $y_{m,d}$, where $d = 1, \ldots, D$, denotes a frequency of the dth term in the vocabulary for the mth document. N_m individual words for the mth document are denoted as $w_{m,n} \in \{1, D\}$, where $n = 1, \ldots, N_m$.
Topic models assume multinomial likelihood

$$\mathcal{L}_m = \prod_{d=1}^{D} q_{m,d}^{y_{m,d}},$$

where $\boldsymbol{q}_m \in \Delta^D$ denotes an unknown expectation parameter of the multinomial distribution, satisfying $q_{m,d} \geq 0$ and $\sum_{d=1}^{D} q_{m,d} = 1$. The expectations $\boldsymbol{q}_m$ decompose as a linear decomposition

$$\boldsymbol{q}_m = \sum_{k=1}^{K} \boldsymbol{\eta}_k \theta_{m,k},$$

where $\boldsymbol{\eta}_k \in \Delta^D$, for $k = 1, \ldots, K$, correspond to **topics** and $\boldsymbol{\theta}_m \in \Delta^K$ to **topic proportions**. Further, assume

$$\boldsymbol{\eta}_k \sim \text{Dirichlet}(\gamma \mathbf{1}), \quad \boldsymbol{\theta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha}).$$

Define an empirical word occurrence distribution

$$p_{m,d} = y_{m,d}/N_m.$$

Inferring $\boldsymbol{q}_m$ that is closer to $\boldsymbol{p}_m$ leads to more accurate topics. The mean multinomial log likelihood relates to the KL-divergence between empirical and latent word distributions,

$$KL(\boldsymbol{p}_m, \boldsymbol{q}_m) = \sum_d \left(p_{m,d} \log p_{m,d} - p_{m,d} \log q_{m,d} \right).$$

The divergence is sensitive to the contribution of *misses*, corresponding to terms for which $\boldsymbol{p}_m$ are *large* but the corresponding $\boldsymbol{q}_m$ are *small*, and, thus, closely relates to the concept of *recall*. Even though, these topics emphasise recall, they may have very low *precision*, containing intruder terms that capture false similarities.

Information retrieval aspect

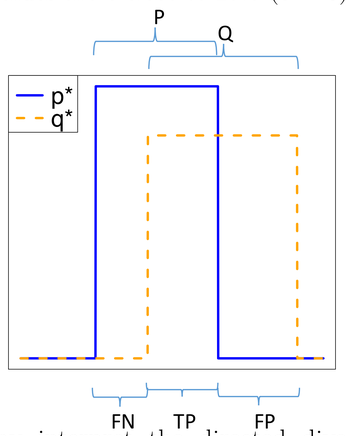
Based on the retrieval model $\boldsymbol{q}_m$, the goal is to retrieve co-occurring terms. Here, the $\boldsymbol{p}_m$ represent relevances (that is, empirical co-occurrences) and $\boldsymbol{q}_m$ should be similar to the $\boldsymbol{p}_m$, avoiding errors. Define two classes of errors, misses and false positives: Terms for which
❶ $\boldsymbol{p}_m$ are large but $\boldsymbol{q}_m$ are small correspond to **misses**, and
❷ $\boldsymbol{q}_m$ are large but $\boldsymbol{p}_m$ are small correspond to **false positives**.
Concepts of recall and precision may be quantified with the directed KL divergences, because $KL(\boldsymbol{p}_m, \boldsymbol{q}_m)$ emphasises misses and the reversed divergence $KL(\boldsymbol{q}_m, \boldsymbol{p}_m)$ emphasises false positives.
Consider maximum entropy distributions for $\boldsymbol{p}$ and $\boldsymbol{q}$ that take uniform values over the support of the distributions, denoted as P and Q, respectively, whereas the remaining values are ϵ-close to zero ($\epsilon \approx 0$).

$$KL(\boldsymbol{p}^*, \boldsymbol{q}^*) = C + \frac{|P \cap Q|}{|P|} \log \epsilon$$

is proportional to *standard* recall, proportion of relevant terms that are retrieved.

$$KL(\boldsymbol{q}^*, \boldsymbol{p}^*) = C + \frac{|P \cap Q|}{|Q|} \log \epsilon$$

is proportional to *standard* precision, proportion of retrieved terms that are relevant.



Because of the connections, we may interpret the directed divergences as generalisations of the concepts of recall and precision for continuously-valued grades of relevances.

Model

Introduce $0 \leq \lambda \leq 1$ that trade-offs the contributions of recall and precision, capturing term co-occurrences and avoiding false similarities. Assume term-specific assignment variables,

$$x_{m,n} \sim \text{Bernoulli}(\lambda), \; c_{m,n} \sim \text{Categorical}(\boldsymbol{\theta}_m).$$

If $x_{m,n} = 0$, with probability $1 - \lambda$,

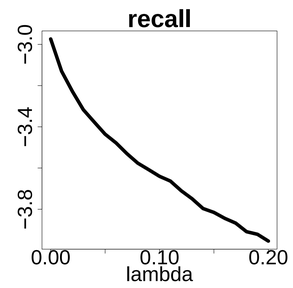
$$w_{m,n} \sim \text{Categorical}(\boldsymbol{\eta}_{c_{m,n}}),$$

otherwise the term is explained by the $\boldsymbol{p}_m$, with probability λ.
Collapsed Gibbs sampling for inference: For $w_{m,n} = d$,

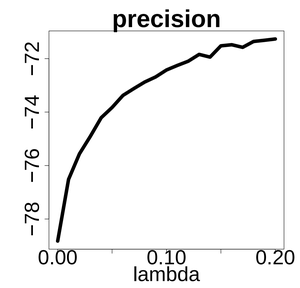
$$p(c_{m,n} = k, x_{m,n} = 0) \propto \frac{N_{k,m}^{-(w_{m,n})} + \alpha_k}{\sum_{k'} N_{k',m}^{-(w_{m,n})} + \sum_{k'} \alpha_{k'}} \times \frac{G_{k,d}^{-(w_{m,n})} + \gamma}{\sum_{d'} G_{k,d'}^{-(w_{m,n})} + \gamma D}$$

$$p(x_{m,n} = 1) \propto \frac{\lambda}{1 - \lambda} p_{m,d}.$$

Results



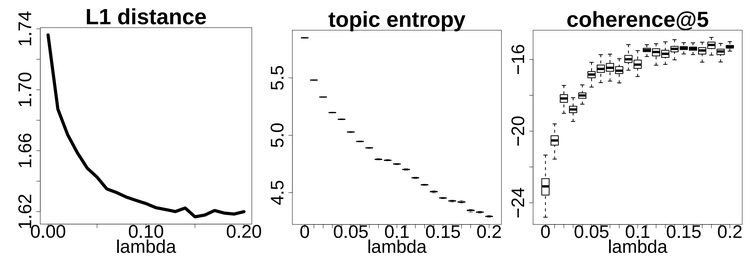
$-\frac{1}{\widehat{M}} \sum_m KL(\widehat{\boldsymbol{p}}_m, \widehat{\boldsymbol{q}}_m)$



$-\frac{1}{\widehat{M}} \sum_m KL(\widehat{\boldsymbol{q}}_m, \widehat{\boldsymbol{p}}_m)$

Precision-recall trade-off:

- Recall is maximal for $\lambda = 0$ and decreases for increasing λ.
- Precision increases for increasing values for λ.



Precision is positively associated with the other performance measures, excluding recall. For increasing λ, and higher precision,

- mean ℓ_1 distances between $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{q}}$ become smaller,
- the topics become more sparse, and
- the semantic coherences for the topics increase.

Acknowledgements

The authors were supported by the EPSRC grant EP/P020720/1, Inference COmputation and Numerics for Insights into Cities (ICONIC), `https://iconicmath.org/`

UNIVERSITY OF CAMBRIDGE