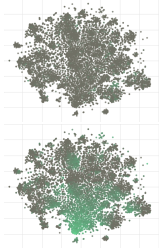
SAP-sLDA: An Interpretable Interface for Exploring Unstructured Text

Charumathi Badrinath, Weiwei Pan, Finale Doshi-Velez

Exploring Text Corpora

Low-dimensional projections are a common way to explore text corpora; ideally, thematically similar documents are clustered together in projected space.

Problem: Popular algorithms for dimensionality reduction of text corpora, like Latent Dirichlet Allocation (LDA), do not produce *useful* projections.



Topic 17 (top) has top 5 words *mindfulness, wisdom, develop, quality* and *concentration*. Topic 36 (bottom) has top 5 words *get, see, let, want* and *know*. More green = higher concentration of that topic.

Contributions

We propose **S**emantically **A**ligned **P**rojection (focused) supervised **LDA** (**SAP-sLDA**) – a flexible LDA-based framework for learning topic models with human-in-the-loop feedback to produce useful low-dimensional projections of corpora. Our method has the explicit objectives of ensuring **semantic alignment** and **robustness** in projections.

We perform a **feasibility analysis** of our method on various synthetic corpora. We also perform preliminary experiments on a **real corpus** from *Dharma Seed*.

The SAP-sLDA Objective

SAP-sLDA objective = LDA ELBO + **regularizer**.

Documents with **same label** pushed **closer** in projected space, documents with **different labels** to pulled **farther**. Projected space = PCA-transformed.

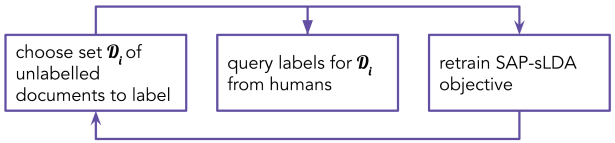
$$\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L}\mathbb{I}(\ell \neq \ell')\cdot\lambda_1\text{dist}(S_\ell, S_{\ell'}, \lambda_2)$$

$$-\mathbb{I}(\ell = \ell')\cdot\lambda_3\text{dist}(S_\ell, S_{\ell'}, \lambda_4)$$

$$\text{dist}(S_\ell, S_{\ell'}, \lambda) = \sum_{x\in S_\ell}\sum_{x'\in S_{\ell'}} ||f(x) - f(x')||_\lambda$$

The SAP-sLDA Algorithm

SAP-sLDA iterates the following steps until projections are stable across random restarts:

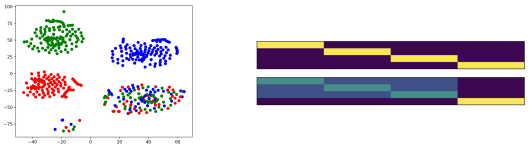


The learned topics for each document are then plotted in 2-D after applying t-SNE.

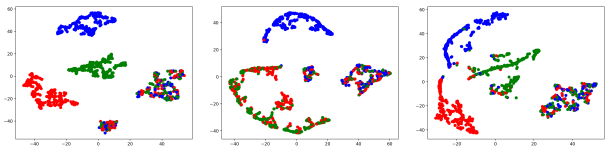
1) How to choose $\mathcal{D}_i$?
2) What sort of **label class** to choose?

Recovering Ground-Truth Data Structure

Synthetic data generated with varying settings of document-topic distribution θ and word-topic distribution β. One example below:



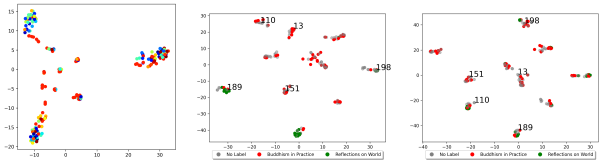
Finding 1: SAP-sLDA consistently **recovers ground truth clusters** while preserving local features for identifiable (left) and non-identifiable β (middle). LDA fails for non-identifiable β (right).

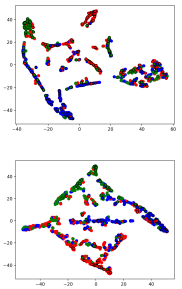


Finding 2: When SAP-sLDA learns a faithful $\hat{\theta}$, the **learned $\hat{\beta}$ is also close to the ground truth**.

Factors Influencing Clustering Quality

Finding 3: The class of labels provided matters. Tested label classes with varying levels of signal: *random*, labelling by *author* (left), labelling by *binary theme* (right two). More semantically-aligned, stable clusters when using binary theme as label class.





Finding 4: Which documents are labelled matters. Tested *random* (top) and *variance-based* (bottom) active learning schemes. Variance-based achieves interpretable clustering with fewer labels (15%) than naive alternative (25%).

Future Work

1) More exploration into label classes; labeling schemes.
2) Tuning hyperparameters and adding further components to regularizer (e.g. topic sparsity; orthogonality).
3) Human studies to investigate whether projection is practically useful for navigating corpus.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1750358. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. A grant from Harvard College URAF's conference funding program is supporting CB's travel costs for attending ICML. We would like to acknowledge the Cambridge Insight Meditation Center and Leandra Tejedor for conducting initial work on the project including dataset collection and exploring data preprocessing techniques.

References

1. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res., 3(null):993–1022, mar2003. ISSN 1532-4435.7
2. Ren, J., Kunes, R., and Doshi-Velez, F. Prediction focused topic models via vocab selection. CoRR, abs/1910.05495, 2019. URL http://arxiv.org/abs/1910.05495.