**PROJECT PROPOSAL: Topic Classification Using Labeled LDA**

Pragya Agarwal, Abhishek Bafna, Kritika Versha

# **Problem Statement:** This project is meant to use credit attribution to verify the pertinence of user assigned labels

**Description:** A significant portion of the world’s text is tagged by readers on social bookmarking websites. Credit-attribution involves, associating each word in the document with most appropriate tag. This is an inherent problem because the multiple tags associated with each document do not occur with equal specificity across the whole document. We plan to address this problem using Labeled LDA a topic model that constrains Latent Dirichlet Allocation

by defining a one-to-one correspondence between LDA’s latent topics and user tags.

**Possible Extension**: Tag-specific snippet extraction using Labeled LDA

**Individual Contribution:**

1. Abhishek: Implementation of Labeled LDA and application to the dataset
2. Kritika: Implementation of Multinomial Naive Bayesian framework and application to dataset.
3. Pragya: Implementation of Multi-Class SVM and application to the dataset
4. Common Tasks : Comparison of the three applied methods and critical analysis of performance, advantages and drawbacks of the three methods.

**Review Paper:** Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

**Dataset:** We plan to use one of the following two large-scale folksonomy data sets were used for the simulation of PINTS. They were obtained by systematically crawling the Flickr and Del.icio.us portals during 2006 and 2007. The crawls were done in the context of the Tagora project. The crawling targets were the core elements, namely users, tags, resources and tag assignments.

The statistics of the crawled datasets are summarized below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Users | Tags | Resources | Tag assignm. | Download |
| Flickr | 319,686 | 1,607,879 | 28,153,045 | 112,900,000 | [flickr\_UsrResTag.7z](http://www.isweb.uni-koblenz.de/files/datasets/pints/flickr_UsrResTag.7z)  (518 MB) packed with 7zip |
| Delicious | 532,924 | 2,481,698 | 17,262,480 | 140,126,586 | [delicious\_UsrResTag.7z](http://www.isweb.uni-koblenz.de/files/datasets/pints/delicious_UsrResTag.7z)  (848 MB) packed with 7zip |

The archives were compressed with 7zip and contain a single text file with time-ordered tag assignments in 4 tab-separated columns. The columns are (in following order): posting date, user ID, resource ID, and tag label.

**Programming Language:** Python