



Overview

The **MITLL Topic Clustering System** performs state-of-the-art topic clustering (i.e. topic-based unsupervised grouping of documents) on a set of text documents after filtering based on language identification. The number of topics automatically extracted from the input documents is a parameter of the system and can be specified by the user. Results are stored in human/machine readable files that can be used for browsing or for integration in further processing stages.

Details

Provided a collection of text documents, the **MITLL Topic Clustering System**:

1. Normalizes the input text and removes non-informative terms
2. Performs language identification on each document
3. Filters out all documents not matching the user-specified language
4. Uses the latent modeling technique called Probabilistic Latent Semantic Analysis (PLSA) to perform a soft classification of the documents into topics by:
 - a. Learning topic classes in an unsupervised fashion
 - b. For each document, assigning a degree of membership to each of the learned topic classes
5. Stores the results in files that can be read by a human for data exploration, or a machine for integration with other applications.

Example

Example topics extracted from a data collection from Kiva (www.kiva.org)

Relevant Terms	
Topic 1	group, lending, member, collateral, guarantee, pressure, repayment, repay, peer, solidarity
Topic 2	milk, farming, dairy, cow, farmer, production, school, poultry, sale, produce
Topic 3	sewing, machine, tailoring, shop, materials, orders, seamstress, day, dresses, makes

Prerequisites

The system is a command-line application for 64-bit Linux written in Python and C⁺⁺. It requires the following dependencies:

- Python 2.7
- NumPy (Python module)

Benefits

Data analysts and developers:

- Quick characterization of large text collection through the discovery of topics
- Soft classification of documents allows users to tag, filter, and group documents based on topic membership in situations where manual labeling of data is unavailable or otherwise infeasible