Regularized Latent Semantic Indexing Source Code

IIP Lab, Nankai Unversity

Introduction

RLSI (Regularized Latent Semantic Indexing) is a non-probabilistic topic model proposed in [1]. The Java package parallelizes RLSI algorithm via multi-threading.

[1] Quan Wang , Jun Xu , Hang Li , Nick Craswell, Regularized latent semantic indexing, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, July 24-28, 2011, Beijing, China

Input:

Term-document matrix

The first line contains the number of terms and documents in the collection.

Each line contains the tf-idf information of term

<term> $t_{c,d}$ t_c *0:tf_idf 1:tf_idf ……*

$\boldsymbol{t_{c,d}}$ document frequency

$\boldsymbol{t_c}$ term frequency in the collection

0:tf_idf tf_idf in document 0

1:tf_idf tf_idf in document 1

……

Output:

a. Term-topic matrix (U)

b. Topic document matrix (V);

Usage:

java RegularizedTopicModel -d word_doc_file_name [options]

Topic model options:

-t int -> number of topics (default 50)

-l1 float -> L1-norm parameter for U (default 0.5)

-l2 float -> L2-norm parameter for V (default 0.5)

Output options:

-v string -> the filename prefix for outputted V matrix (default VMatrix)

-u string -> the filename prefix for outputted U matrix (default UMatrix)

-sv int -> number of skipped iterations that don't output V (default 5)

-su int -> number of skipped iterations that don't output U (default 5)

Optimization options:

-# int -> number of learning iterations (default 500)

-c int -> number of threads running in parallel (default 4)

Restart options:

-restart v -> restart from initial V matrix

-iv string -> filename of initial V matrix

Development Team:

Zhicheng He, Yingjie Xu, Jun Xu, MaoQiang Xie, Yalou Huang

Maintained by Lab of Intelligent Information Processing, Nankai University, China.

(Last Update: Mar. 26, 2013)