

R によるステミング

■ lsa(CRAN)

- Latent Semantic Analysis <http://cran.r-project.org/web/packages/lsa/index.html>

The basic idea of latent semantic analysis (LSA) is, that text do have a higher order (=latent semantic) structure which, however, is obscured by word usage (e.g. through the use of synonyms or polysemy). By using conceptual indices that are derived statistically via a truncated singular value decomposition (a two-mode factor analysis) over a given document-term matrix, this variability problem can be overcome.

■ RWeka

- R/Weka interface <http://cran.r-project.org/web/packages/RWeka/index.html>

An R interface to Weka (Version 3.7.1). Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Both the R interface and Weka itself are contained in the RWeka package. For more information on Weka see <http://www.cs.waikato.ac.nz/~ml/weka/>.

■ Snowball

- Snowball Stemmers <http://cran.r-project.org/web/packages/Snowball/>

Snowball stemmers.

■ tm

- Text Mining Package <http://cran.r-project.org/web/packages/tm/index.html>

A framework for text mining applications within R.

R 以外によるステミング

■ Lingua::Stem(CPAN)

- <http://snowhare.com/utilities/modules/lingua-stem/>
- Lingua::Stem::En
 - <http://search.cpan.org/~snowhare/Lingua-Stem-0.84/lib/Lingua/Stem/En.pm>

■ 参考資料

- <http://geek.blog.eonet.jp/memo/2009/04/index.html>