

A Text Mining Framework in R and Its Applications

Ingo Feinerer

Department of Statistics and Mathematics,
Wirtschaftsuniversität Wien, Austria

defensio dissertationis, 22.10.2008



Motivation

- ▶ Vast amount of textual data available in machine readable format:
 - ▶ scientific articles,
 - ▶ abstracts,
 - ▶ books,
 - ▶ memos,
 - ▶ letters,
 - ▶ online forums,
 - ▶ mailing lists,
 - ▶ blogs,
 - ▶ ...
- ▶ Steady increase of text mining methods (both in academia as in industry) within the last decade

Text Mining

- ▶ Highly interdisciplinary research field utilizing techniques from computer science, linguistics, and statistics
- ▶ Automated processing of text (Miller, 2005)
- ▶ A knowledge-intensive process (Feldman and Sanger, 2007)
- ▶ Use of text collections to discover new facts (Hearst, 1999)
- ▶ Analyzing unstructured information (Weiss et al., 2004)
- ▶ Intelligent text processing (Gelbukh, 2004)

Existing Tools and Infrastructure

- ▶ Several commercial text mining products exist, but
 - ▶ you do not know how methods really work (closed source)
 - ▶ it is hard to extend the tools with your own features (poor API functionality, different plug-in mechanisms)
 - ▶ they are very expensive
- ▶ On the other hand, we have an excellent free and open source software environment for statistical computing: R
 - ▶ provides a broad range of methods for clustering, classification, visualization (i.e., the algorithmic foundations for text mining)
 - ▶ is used to implement research prototypes (so you have state-of-the-art algorithms typically not available in commercial products)
 - ▶ however lacked an explicit infrastructure to apply all this functionality on texts (until now!)

Research Aims

- ▶ Development of an open source text mining infrastructure for R such that the integration with existing functionality in R offers capabilities similar to established commercial text mining tool kits
- ▶ Evaluation and validation of the text mining infrastructure via applications to real data from various fields, like e-commerce and law

Text Mining Package and Infrastructure



I. Feinerer

tm: Text Mining Package, 2008

URL <http://CRAN.R-project.org/package=tm>

R package version 0.3-2



I. Feinerer, K. Hornik, and D. Meyer

Text mining infrastructure in R

Journal of Statistical Software, 25(5):1–54, March 2008

ISSN 1548-7660

URL <http://www.jstatsoft.org/v25/i05>

Conceptual Layers and Packages

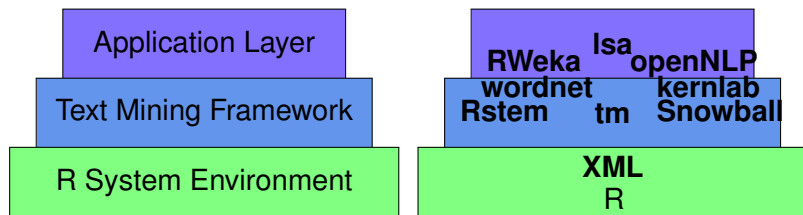


Figure: Conceptual Layers and Packages.

UML Class Diagram

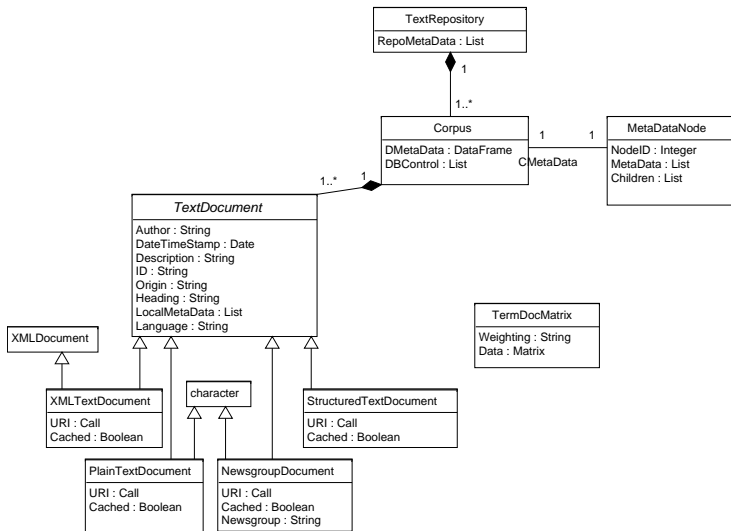


Figure: UML class diagram of the **tm** package.

Sources

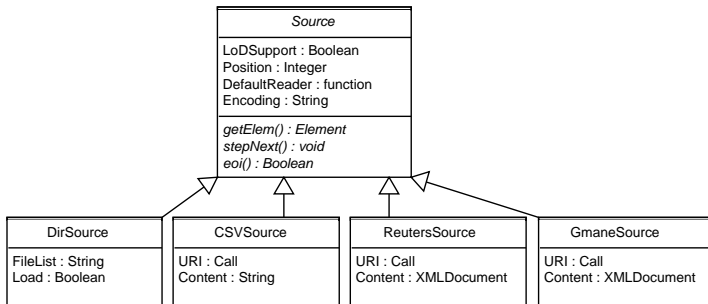


Figure: UML class diagram for Sources in the **tm** package.

Algorithms

- ▶ Document and Corpora Handling:
 - ▶ Constructors
 - ▶ Merging
 - ▶ Accessors and Extractors
- ▶ Transformations: Define mappings for corpora
 - ▶ Most preprocessing functions are transformations
 - ▶ Capture the concept of maps from functional programming
- ▶ Filters: Define predicate functions to extract documents from corpora
 - ▶ Full text search
 - ▶ Filters have full access to meta data

Extensions

- ▶ Modular structure designed for easy extensibility
- ▶ Readers, sources, etc. define interfaces
- ▶ Implementations for these interfaces generate first-class objects (internal objects use the same mechanism)
- ▶ Easy plug-ins, we provide the infrastructure, the (advanced) user his custom functionality

Predefined Functionality

However, most users want to use **tm** out of the box.

- ▶ Preprocessing: data import, stemming, stopword removal, part of speech tagging, synonyms, . . .
- ▶ Basis analysis techniques: count based evaluation, text clustering, text classification, . . .
- ▶ Access to more advanced functionality: full integration with string kernels, latent semantic analysis, . . .
- ▶ Export of term-document matrices: basically all methods in R working on matrices

Text Mining for B2C E-Commerce

Use text mining to find out facts on

- ▶ product development,
- ▶ product improvement, and
- ▶ consumer feedback.

Analyze texts dealing with customers' opinions, requests, complaints, ideas and questions.

We performed an analysis of the support forum of a major content management system.

Text Mining for B2C E-Commerce

Results

We could find potentially business relevant information via

- ▶ considering the most frequent terms
- ▶ associations between different constructs
- ▶ grouping into classes of interest
- ▶ supervised clustering (i.e., help out with a priori known classifications)
- ▶ extracting relations to competitive products
- ▶ churn analysis based on negative terms

Law Mining

Text Mining of Austrian Supreme Administrative Court Jurisdictions

- ▶ Jurisdictions need to be clustered into tax classes
- ▶ Experts for Austrian tax law were especially interested in:
 - ▶ Verfahrensanlässe
 - ▶ Behördentyp
 - ▶ Aufhebungsgründe
 - ▶ Entscheidungsstruktur der Senate
 - ▶ Verfahrensdauer
- ▶ I.e., we want to find patterns in jurisdiction

Law Mining

Results



I. Feinerer and K. Hornik

Text mining of supreme administrative court jurisdictions
Data Analysis, Machine Learning, and Applications
Studies in Classification, Data Analysis, and Knowledge
Organization. Springer-Verlag, 2007.

- ▶ k -means clustering and keyword based clustering work well enough to reasonably group documents for further investigation by law experts
- ▶ Deriving the senate size works very well
- ▶ We presented approaches that can clearly aid legal experts to process their documents faster and more focused

Wizard of Oz Stylometry



I. Feinerer

An introduction to text mining in R

R News, 8(2):19–22, October 2008.

URL <http://CRAN.R-project.org/doc/Rnews/>

- ▶ Can we distinguish the two main authors of the Wizard of Oz book series (Baum and Thompson)?
- ▶ It is a known hard problem, there has been a long dispute between literature experts on the 15th book of Oz
- ▶ Extraction of relevant terms and principal component analysis show weak support for Thompson as the author of the 15th book
- ▶ This classification is now commonly accepted, however our approach can only indicate the direction, not significantly distinguish

Who is actually using **tm**?

- ▶ **tm** is now *the* official text mining infrastructure of R (and from the amounts of e-mails I receive we know it is actually used)
- ▶ **tm** was used for analyses by various researchers resulting in publications (so we know the usage from the references)
- ▶ **tm** is known to be deployed in a social science course at the Justus-Liebig-Universität Gießen (Germany)
- ▶ **tm** was officially evaluated by a governmental employee of the Australian taxation office
- ▶ **tm** was recently considered as the basis for a natural language processing tool kit by Harvard University (USA) (talks about possible cooperation and synergy effects in progress)

Outlook

tm provides a well-grounded foundation, however there is always room for improvement:

- ▶ The package could be faster. The situation will improve within the next releases of R where better character handling (caching) has been promised.
- ▶ Provide methods for sparse computation (necessary for large data sets/industrial scenarios). Work in progress, e.g., for sparse spherical k -means clustering.
- ▶ **tm** provides the text mining infrastructure but users want polished all-in-one solutions. I.e., special application layers implementing typical process structures will be a prerequisite for broader application of **tm** in business intelligence.

Conclusion

- ▶ Presentation of a modular, extensible, thoroughly designed text mining infrastructure for R
- ▶ Application of **tm** in diverse fields, like law, e-commerce or stylometry
- ▶ We could show that **tm** is now well established in academia and has also been applied for real analysis in various areas of interest