etea: Classifier-leverage functions for knowledge discovery and predictive modelling from unstructured texts in 'R'

An extended abstract to outline concepts

Christopher S Kirk, PhD, FIAP
Bristol, UK

Overview

In order to analyse events that involve human behaviour or the interactions between human and organisation in diverse fields such as in special investigations, clinical treatment systems or capital market transactions, investigators benefit from examination of the content of clinical, case or transaction notes and records. Examination typically comprises the application of methodologies such as feature-extraction, classification and 'sentiment' valence analysis. By pre-processing unstructured text into components that also group transitions together, the potential for added value derived from adaptive algorithms processing and post-processing can be leveraged significantly to provide a multi-dimensional classification of textual content over state or time. The majority of cases can be described using classification categories grouped into Environment, Thought, Emotion and Action, (etea) from which the title of this package has been derived. Notably, these groupings also promote temporal analysis.

This 'R' (1), package is expected to be of interest to all those who seek to enhance knowledge discovery and to model using unstructured text as the data source.

Typical cases include:

behavioural analysis and segmentation studies of markets,
scoring of texts for sequencing models or signature channels
special interest groups and investigations:
Families with Complex Needs,
Special Victims Units and Victim Support,
Clinical or Treatment Systems

Recent work

In pursuing investigations that consider causality, traditional sentiment analysis techniques have been augmented (2), to provide a structure that is refined enough for use in post-classification 'text-emotion' modelling. Parallel, state-space work exploring partially observable, markovian decision problems found that complex behaviours, particularly those involving multiple agents, are computationally expensive and can only be solvable by integrating multiple events into a set of primitive states. However, whilst doing so permits rationalisation of large-scale problems and the successful evaluation of states, actions and observations, care must be taken not to lose the segmentation of text-notes offered by state Ids or date-time-stamps. The drive to elaborate a solution inspired the creation of the leverage-functions described here.

During an evaluation of multi-state modelling to describe event progression, it was concluded that the addition of a dimension(s) to model evolution that occurred within each state would be useful. To further this work, text-mining of case, clinical and transaction textual notes has been examined and shows promise by providing a temporal and state condition framework that suggests it might enhance predictive potential. The discretisation of textual-features into transition or domain intervals, (Table 1) to add the dimensions of environment, thoughts, emotions and actions, (etea) has been trialled using test data. In order to examine the hypothesis that a quantifiable, progressive relationship exists between the environment, thoughts, emotions and actions that is identifiable to the extent that reliable predictions of future action (and therefore environmental change) can occur, it was necessary to elaborate functions to promote algorithmic development.

Possible use

Given this test success, it is suggested that it may now be possible to mathematically map the potential for the evolution of an action occurring (observable) as a consequence of thought and emotions developing, (unobservable) and driven by the micro-environment (observable) and this will undoubtedly be of use to those who investigate corporate or human behaviour and who are able to deploy avoidance measures through early intervention or negate the need for reparation measures. It is acknowledged of course, that unstructured text is often captured within a skewed set, namely those who are already in a system (such as in a single intelligence stream, treatment system or special interest/care group) and that extrapolation to those outside the studied system will need caution. However, by modelling text or data from more than one system such as that derived from related intelligence or source feeds, the neural network/state-space model becomes a deep network model with a composition of multiple non-linear layers when unfolded in time.

Table 1: Domain/Transition Intervals for Special Interest Groups

Environment	**Thought**	**Emotion**	**Action**
Observable	***Unobservable***	***Unobservable***	***Observable***
Built Environment	Analysis	Anger	Communicate
Relations and Associations	Application	Disgust	Create
Money and Food	Comprehension	Fear	Handle Data
Mental and Physical Health	Contemplation	Joy	Help
Welfare and Well-being	Knowledge	Sadness	Manage or Organise
	Planning	Surprise	Research or Teach
			Technical
			Specialist Interest

Suggested methodology

The classification function is designed to maximise flexibility by enabling the deployment of a user-defined lexicon of classification terms and categories. This functionality promotes adaptation to meet specific user-need in specialist cases and updating the lexicon is simplified by a supplied helper function. To exploit this flexibility, the user might typically create a text-note extraction file of particular interest and then add state Ids and a synthetic date-time difference of lapsed time. Examples are the study of risk, human exploitation, treatment recovery and capital markets transactions with a 'date difference' (for each text-note) of days elapsed from a baseline start say the first date from a group of clients' data.

These extraction files are then the focus of a text-mining, knowledge-discovery function, (create_q_matrix.R) that calls functions (3), to build a frequency matrix (primary pre-processing and classification) of features present in each sentence of the document/note under investigation. This creates added-value for investigators and can be used as a basis for further work including updating the user-defined lexicon noted above. Notably, the function decomposes each document text-note into sentences so that the output of further processing is not averaged which would dilute the value of the classification. Optionally, the output matrix includes row-names that comprise state and/or time-stamp identification to benefit further processing. This ensures that the integrity of the state/time ID is not lost when sentences are created from timestamped text-notes. The function, (etea_classify.R) uses this aforementioned matrix ('tm' (4) or 'quanteda' format) and by calling the user-defined reference lexicon of words of interest (including categories) quantifies the word frequencies per category and identifies the valence polarity of the sentence. This output can then be used to further quantify frequencies to create alerts or as input into other functions for other processing such as neural networks, state-space modelling or deep network aggregation. This secondary pre-processing and classification has a significant influence on the potential for predictive modelling and post-processing.

Elaborated functions

To broaden the range of this methodology (and potential) for other investigators, the author has elaborated these functions in the 'R' programming language, published them as the 'R' package named 'etea' and to improve compatibility with several other open source software licenses, licensed it with a free software license, GPLv3.

R is a language and environment for statistical computing and graphics. R is designed as a true computer language with control-flow constructions for iteration and alternation, and it allows users to add additional functionality by defining new functions.

Location:	https://github.com/chriskirkhub/etea
Maintainer:	drchriskirk@gmail.com

References

1. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
2. sentiment: Jurka, T; An R package with tools for sentiment analysis including bayesian classifiers for positivity/negativity and emotion classification. URL https://github.com/timjurka/sentiment
3. quanteda: Benoit, K. and Nulty, P.; A fast, flexible toolset for the management, processing, and quantitative analysis of textual data in R. URL https://cran.r-project.org/web/packages/quanteda/
4. tm: Feinerer, I; A framework for text mining applications within R. URL https://cran.r-project.org/web/packages/tm/

Acknowledgements

The functions described herein are based, with permission, on an original methodology for classification of emotions elaborated by Tim Jurka.
The functions described herein call functions, with permission, from the 'quanteda' package elaborated by Ken Benoit and Paul Nulty.
The author expresses thanks for such permissions being granted and for kind assistance throughout.

Document Version Control

Document Version Number 2
Dated June 2016