etea: classifier leverage ('R') functions for predictive modelling from unstructured texts

An extended abstract outlining concepts on the release of the 'R' package: 'etea'

Christopher S Kirk
Bristol, UK

Overview
In order to investigate events that involve human behaviour or the interactions between human and organisation such as in clinical treatment systems or capital market transactions, investigators need to examine the content of clinical, case or transaction notes and records. Such examination is possible by the application of methodologies such as feature-extraction, classification and the so-called 'sentiment' valence analysis. These can quantify the attraction or aversion to an event or situation such as the study of the emotional content of documents. The work described in this paper enables the classification of unstructured textual data into a structured, segmented, temporal frequency matrix for use as input into predictive modelling systems such as neural networks or state-space models.

Recent work
In pursuing investigations that consider causality the valence analysis techniques noted above can be lacking in providing a structure that is refined enough (and complete enough) for use in predictive modelling. Parallel 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. This permitted rationalisation of extensive problems and permitted evaluation of states, actions and observations.

During an evaluation of multi-state modelling to describe events it was concluded that the addition of a dimension(s) to model evolution 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. Therefore discretisation of textual-features into 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 ease algorithmic development.

Possible use
Given test success it is suggested that it may 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 or negate the need for reparation measures.

Table 1: Intervals/Data Reduction Primitives

Environment	Thought	Emotion	Action
Structure	Analysis	Anger	Communicate
Money/Food	Application	Disgust	Create
Health	Comprehension	Fear	Data
Welfare	Contemplate	Joy	Help
	Knowledge	Sadness	Manage/Organise
		Surprise	Research/Teach
			Technical
			Specialist Interest

Suggested methodology
The large body of textual data is extracted and segmented to isolate sectors of particular interest. Examples are the study of human exploitation, treatment recovery and capital markets transactions.

These extraction files are then made the subject of a text-mining knowledge-discovery function that builds a frequency matrix (primary pre-processing and classification) of features present in each sentence of the document (note) in the investigation. This creates an added-value for investigators and can be used as a basis for further work. However, these findings can be further discretised by splitting the continuous range (of a corpus), again by document, into intervals, (etea). 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 two 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