

Using Regular Expressions and Optimisation to Create a Method for Automatically Extracting Dates from War Diaries

The digitisation and transcription of collections of historic text has allowed researchers to use natural language processing techniques to analyse large volumes of this data. Of particular interest is the use of natural language processing to analyse historic diaries. An advantage of analysing diaries is that they contain temporal information regarding the sequence the text was written in. However, dates within such diaries are not always written in a clear and consistent format. Due to this, extracting this temporal information can be quite challenging and time consuming. Using a collection of Australian World War I diaries held by the State Library of New South Wales as a case study, we have devised a method for automatically extracting these dates.

Initially dates are extracted using a variety of regular expressions. From these extracted dates several problems are noticed. Some dates will be missing information as a diarist may have decided it was not always necessary to write the month and year. Furthermore, dates were sometimes written incorrectly due to confusion whilst on the front line. Also, whilst each entry will have a date, dates can also be mentioned within the entry. It is necessary to differentiate between these entry and non-entry dates. Furthermore, due to the wide variety of regular expressions used to capture all dates, non-dates are sometimes extracted and these need to be removed. To address these problems an optimisation program is used to turn the raw dates into a well formatted list of the correct entry dates. This program can also highlight issues such as the diary being written out-of-order.

This talk will explore this date extraction method, in particular why the optimisation program was formatted in the way it was. Furthermore, the accuracy of this method will be discussed, including the use of a GUI written using PySimpleGUI to manually create a set of testing data.