# Wikipediabase Date Parser

Chris Perivolaropoulos

Sunday 21 February 2016

## Contents

Dateparser resides in a separate package called overlay-parse

## 1 Parsing with overlays

The concept of an overlay was inspired by emacs overlays. They are objects that specify the behavior of a subset of a text, by assigning properties to it. An overlay over a text $t$ in our context is tuple of the range within that text, a set of tags that define semantic sets that the said substring is a member of, and arbitrary information (of type $A$) that the underlying text describes. More formally:

$$o_i \in TextRanget \times Set(Tag) \times A$$
$$Text \rightarrow \{o_1, o_2, ..., o_n\}$$

So for example out of the text

$$The weather today, \overbrace{Tuesday}^{o_1} \overbrace{21^{st}}^{o_2}\ of\ \overbrace{November}^{o_3} \overbrace{2016}^{o_4}, was\, sunny.$$

We can extract overlays $\{o_1, ..., o_4\}$, so that

$$\begin{array}{llll} o_1 = ( & r("Tuesday"), & \{\text{DayOfWeek}, \text{FullName}\}, & 2) \\ o_2 = ( & r("21^{st}"), & \{\text{DayOfMonth}, \text{Numeric}\}, & 21) \\ o_3 = ( & r("November"), & \{\text{Month}, \text{FullName}\}, & 11) \\ o_4 = ( & r("2016"), & \{\text{Year}, \text{4digit}\}, & 2016) \end{array}$$

Notice how for all overlays of the example we have $A = \mathbb{N}$, as we encode day of the week, day of the month, month and year as natural numbers. We encode more precise type information (ie that a day is inherently different than a month) in the tag set.

Once we have a set of overlays we can define overlay sequences as overlays whose ranges are consecutive, that is their and their tag sets match particular patterns. For example we can search for sequences of overlays that match the pattern

$$p = \text{DayOfMonth}, \text{Separator}(/), (\text{Month} \wedge \text{Number}), \text{Separator}(/), \text{Year}$$

to match patterns like 22/07/1991, where $Separator(/)$ matches only the character "/"

# 2 The implementation

# 3 Optimization

## 3.1 Comparison

# 4 The dates example

# 5 Benchmarks