Bis zum Sankt(-|\s)?[Nn]immerleins(-|\s)?[Tt]ag

DER DATUMSERKENNER »PDR-DATES«

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Martin Fechner (TELOTA), Fabian Körner (PDR)

AUFGABE

Im Sinne eines Data Retrieval ist es sinnvoll Datumsangaben in heterogenen, verschiedensprachigen Texten mit maschinenlesbarem Markup zu versehen. Bei großen Datenmengen wird bei manuellem Vorgehen für die Identifizierung und Kennzeichnung von Datumsangaben ein erheblicher Zeitaufwand benötigt. Die hier vorgestellte Softwarelösung kann Datumsangaben als Zeitpunkte oder Zeiträume automatisch erkennen und wandelt sie in ein Standardformat nach ISO 8601 um.

IDEE

PDR-DATES ist eine Java-Bibliothek, die auf der syntaktischen Mustererkennung durch reguläre Ausdrücke aufbaut. Um komplexere Zeitangaben in Texten erkennen zu können, werden drei Schritte angewandt: (1) Im tokenisierten Text wird mit regulären Ausdrücken geprüft, ob die einzelnen Tokens für eine Datumsangabe relevante Informationen enthalten. (2) Über mehrere klassifizierte Tokens hinweg wird nach definierten Mustern gesucht. Diese Mustererkennung wird mit einer Vielzahl von Mustern über dem gleichen Text wiederholt, so dass auch zusammengesetzte Datumsangaben erkannt werden können. (3) Schließlich werden alle erkannten Datumsangaben hinsichtlich ihrer Bedeutung interpretiert. Das macht auch die Erkennung von festen und beweglichen Feiertagen möglich.

KONFIGURATION

Die bereitgestellte Bibliothek erkennt Datumsangaben in deutsch, englisch und italienisch. Mit Hilfe einer XML-Konfigurationsdatei ist es darüber hinaus möglich, eine eigene Java-Bibliothek zu erzeugen und PDR-DATES so um eigene Ausdrücke zu erweitern. Damit kann die Datumserkennung an einzelne Forschungskontexte angepasst und zwischen möglichen Kooperationspartnern ausgetauscht werden.

AUSBLICK

Mit dem geschilderten Vorgehen werden ausschließlich vollständige Datumsangaben erkannt. Für eine Interpretation von Datumsangaben, die sich nur relativ zu einem Bezugsdatum interpretieren lassen (etwa: "letzte Woche"), ist es denkbar, die syntaktische Mustererkennung auch um eine semantische Mustererkennung zu erweitern.

ANWENDUNGEN

Der Datumserkenner ist als Webservice frei abrufbar und kann die Auszeichnung von Texten mit Markup unterstützen und die Eingabe von Datumsangaben in Webformularen (etwa in correspSearch) erleichtern.

```
<config>
    <language>De</language>
    <yearspan>
        <start>1582</start>
        <end>2100</end>
    </yearspan>
    <!-- REGULÄRE AUSDRÜCKE -->
    <constant id="year">
        <regEx>\\d\\d\\d\\d</regEx>
        <rend type="other" label="YEAR"/>
    </constant>
    ...
    <constant id="month_02">
        <regEx>Feb|Februar</regEx>
        <rend type="date" label="MONTH">
            <date>
                <month>02</month>
            </date>
        </rend>
    </constant>
    ...
    <group id="month" label="MONTH">
        ...
        <ref id="month_02"/>
        ...
    </group>
    ...
    <!-- MUSTER -->
    ...
    <pattern id="month_yyyy" type="Date">
        <symbol id="month"/>
        <symbol id="space"/>
        <symbol id="year"/>
    </pattern>
    ...
</config>
```

BEISPIEL

Ab Mitte Februar bis Ostern 2016 ist ein kurzer Zeitraum, ab etwa Juli 2009 bis etwa 2015 ein längerer.

(1) Reguläre Ausdrücke unterteilen die Token in einzelne Klassen, so werden »Anfang«, »Januar« und »2016« als »approximation«, »month01« und »d4« erkannt. Neben Zahlen und Zahlausdrücken werden Feiertage, Monatsnamen, Jahreszeiten, Näherungsangaben, Jahrhunderte und Wörter mit Sonderfunktion identifiziert:

Text:	ab	etwa	Juli	2009	bis	etwa	2015	ein	längerer	.
Code:	LIMIT	APPROX	month	YYYY	LIMIT	APPROX	YYYY	WORD	WORD	PUNCT
	(from)	(circa)	(7)	(year)	(to)	(circa)	(year)			

Text:	Ab	Mitte	Februar	bis	Ostern	2016	ist	ein	kurzer	Zeitraum	,
Code:	LIMIT	APPROX		TO	HOLYDAY	YYYY	WORD	WORD	WORD	WORD	PUNCT
	(from)	(2/3)	(2)	(to)	(easter0)	(year)					

(2) Die Mustererkennung gibt den einzelnen Tokens eine vorläufige Bedeutung für die spätere Interpretation. Da im Text iterativ nach verschiedenen Mustern gesucht wird, ist es möglich schon erkannte Datumsangaben durch Konkretisierungen in Form von Prefix- oder Suffix-Mustern zu erweitern. Jede so erkannte Datumsangabe bezeichnet entweder einen Zeitpunkt oder einen Zeitraum:

	Ab	Mitte	Februar bis	Ostern 2016
1				holiday_yyyy
2			month_to_	
3		approx_		
4	limit_			

	ab	etwa	Juli 2009	bis	etwa	2015
1			month_yyyy			yyyy
3		approx_			approx_	
4	limit_					
5				_to_		

(3) Durch Interpretation der Muster wird eine Datumsangabe im Format nach ISO 8601 erzeugt. Einige Token erhalten je nach Positionierung im Muster eine andere Interpretation, so bezeichnet der Term »Anfang« in »Anfang März« und »Anfang 1800« jeweils unterschiedlich lange Zeiträume. Auch können Ausdrücke wie »Mariä Empfängnis« interpretiert werden:

Result 1: „Ab Mitte Februar bis Ostern 2016“
→ notBefore="2016-02-11" to="2016-03-27"

Result 2: „ab etwa Juli 2009 bis etwa 2015“
→ notBefore="2009-05" notAfter="2017"

PDR-Webservice:
https://pdrprod.bbaw.de/pdrws/dates?doc=api

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN