„Bis zum Sankt(-|\s)? [Nn]immerleins(-|\s)?[Tt]ag“ – Der Datumserkenner „PDR-Dates“

,
fechner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

,
fkoerner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Einleitung

Die Idee für einen Datumserkenner „PDR-Dates“ entwickelte sich 2009 während des Anfangsstadiums des von der DFG geförderten Personendaten-Repositoriums an der Berlin-Brandenburgischen Akademie der Wissenschaften . Im Zuge dieses Projektes wurde eine Lösung für die gemeinsame Speicherung und Bereitstellung von Informationen aus heterogenen historischen Personendatenbeständen geschaffen. Diese Bestände stammen vorrangig aus laufenden und abgeschlossenen Projekten der BBAW.

Neben der zu erstellenden Software-Umgebung lag also von Anfang an ein wesentlicher Anteil der Arbeit des PDR im Bereich der Migration bestehender Informationsmengen aus ihrem wie auch immer gearteten Ausgangsformat in das Format des PDR.

Dabei fiel auf, dass man bei dieser Gelegenheit zwei Fliegen mit einer Klappe schlagen und versuchen kann, die Strukturierung der Informationen nicht nur zu übernehmen, sondern zu verbessern. Dabei ist extrem hilfreich, dass man den genauen Kontext der Inhalte einer Informationsmenge nutzen kann, um das global extrem komplexe Problem der Erkennung von Datumsangaben soweit zu reduzieren, das ein realisierbares Werkzeug für die Automatisierung geschaffen werden kann: der Datumserkenner „PDR-Dates“.

Funktionsweise

Ziel der Datumserkenners „ PDR-Dates“ ist die Identifizierung von natürlich-sprachlichen Datumsangaben in verschieden-sprachigen Texten, um im Sinne des Data Retrieval einen Mehrwert zu erzielen. Natürlich-sprachliche Datumsangaben sollen hierfür in das Standardformat nach ISO 8601 (s. International Organization of Standardization und Wikipedia) umgewandelt werden, damit sie maschinenlesbar sind. Zu diesem Zweck wurde eine Java-Bibliothek programmiert, die in Forschungsumgebungen oder in Web Services integriert werden kann. Beispiele hierfür sind die Zeitraumangaben des Webservices „ corrospSearch “ oder die Webservices des PDR , die mit „PDR-Dates“ arbeiten. Der Datumserkenner „PDR-Dates“ kann sowohl einzelne Zeitpunkte, als auch Zeiträume erkennen.

Der Datumserkenner baut auf der syntaktischen Mustererkennung durch reguläre Ausdrücke auf. Um komplexere Zeitangaben in Texten erkennen zu können, werden drei Schritte angewandt: (1) Der tokenisierte Text wird mit regulären Ausdrücken geprüft, ob die einzelnen Tokens für eine Datumsangabe relevante Informationen enthalten können. (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 lange zusammengesetzte Datumsangaben erkannt werden können. (3) Schließlich werden alle erkannten Datumsangaben hinsichtlich ihrer Bedeutung interpretiert. Dafür wird auf alle zusammengetragenen Informationen zurückgegriffen.

(1) Die regulären Ausdrücke unterteilen die Tokens in einzelne Klassen, so wird: "Anfang", "Januar", "2016" als "approximation", "month01", "d4" erkannt. Neben Zahlen und Zahlausdrücken werden Feiertage, Monatsnamen, Jahreszeiten, Näherungsangaben, Jahrhundertangaben und Wörter mit Sonderfunktion erkannt.

(2) Die Mustererkennung gibt den einzelnen Tokens eine vorläufige Bedeutung für die spätere Interpretation (etwa "März 1800" als "month_yyyy"). 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 ("1.1.1800" als "d_m_yyyy") oder es ist möglich, dass durch ungefähre Angaben ein Zeitraum bezeichnet wird ("Anfang März 1800" als "approximation_month_yyyy"). Auch kann erkannt werden, ob zwei schon erkannte Datumsangaben einen Zeitraum bezeichnen ("von Dezember 1800 bis Januar 1805" als "limit_month_yyyy_to_month_yyyy").

(3) Bei der Interpretation der Muster werden alle festgestellten Informationen für die Verarbeitung zum Format nach ISO 8601 genutzt. Einige Tokens 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 feste Feiertage ("Mariä Empfängnis 1800"), sowie von Ostern und dem Jahr abhängige Feiertage ("Pfingstmontag 1800") interpretiert werden. Handelt es sich bei einem oder beiden Daten bereits um eine Zeitspanne, wird der volle Zeitraum als Zieldatum ausgegeben.

(Beispiel über den Web Service des PRD mit dem Text „Die nächsten Semesterferien dauern von Mitte Februar bis Mitte April 2016“ findet sich hier:
https://pdrprod.bbaw.de/pdrws/dates?lang=de&text=Die%20n%C3%A4chsten%20Semesterferien%20dauern%20von%20Mitte%20Februar%20bis%20Mitte%20April%202016&output=xml)

Erweiterungen und Begrenzungen

Mit Hilfe einer Konfigurationsdatei in XML ist es möglich eine eigene Java-Bibliothek zu erzeugen. Damit kann die Datumserkennung an einzelne Forschungskontexte angepasst und dort zwischen möglichen Kooperationspartnern ausgetauscht werden.

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"), müsste die syntaktische Mustererkennung auch um eine semantische Mustererkennung erweitert werden. Die bereitgestellte Bibliothek (zu erreichen über den Web Service des PDR, für die APIs https://pdrprod.bbaw.de/pdrws/dates?doc=api) erkennt Datumsangaben in deutsch, englisch und italienisch. Mit der geschilderten Konfigurationsdatei ist eine Erweiterung aber auch ohne Programmierkenntnisse möglich.

Bibliographie

correspSearch (2015): *correspSearch – Verzeichnisse von Briefeditionen durchsuchen*. http://correspsearch.bbaw.de/search.xql?l=de [letzter Zugriff 15. Oktober 2015].

International Organization for Standardization (o. J.): *Date and time format – ISO 8601*. http://www.iso.org/iso/iso8601 [letzter Zugriff 15. Oktober 2015].

Personendaten-Repositorium (2012): *Webservices – Personendaten-Repositorium*. http://pdr.bbaw.de/software/webservices/ [letzter Zugriff: 15. Oktober 2015].

Wikipedia: *ISO 8601*. https://de.wikipedia.org/wiki/ISO_8601 [letzter Zugriff 15. Oktober 2015].