GESStabs

Encoding

Gesellschaft für Software
in der Sozialforschung mbH

Waterloohain 6 - 8
22769 Hamburg
Tel.: 040 - 853 753 - 0
Fax: 040 - 853 753 - 33
www.gessgroup.de

Encoding von Inputfiles

Seit dem Frühjahr 2014 verwendet GESStabs intern nur noch den UTF8 Code. Das bedeutet, dass alle Eingabedateien, während der Interpretation dahingehend gewandelt werden. Hierfür muss die Enkodierung jedes Script-, Daten oder OPENQ-Files richtig erkannt werden.

GESS kennt

UTF16LE
UTF16BE
UTF8BOM
UTF8 ohne BOM
Latin1.

Am sichersten fährt man also bei Scriptdateien, wenn man sie UTF8-kodiert speichert, UTF8 deckt (fast) alle darstellbaren Zeichen ab, und man ist auf der sicheren Seite. GESStabs erkennt das Encoding sehr sicher: Unicode (UTF16LE und UTF16BE) wird anhand des BOM (Byte-Order-Mark) erkannt. Dasselbe gilt für UTF8 mit BOM.

Findet GESStabs eine Datei ohne BOM, reduziert sich die Entscheidung auf die Alternative „UTF8 oder nicht“. GESStabs verwendet hier eine sehr einfache Heuristik: wird kein BOM gefunden, wird die gesamte Datei durchforstet, ob sich Zeichen finden lassen, die der Bildungslogik von UTF8 Zeichen widersprechen. Sind alle Zeichen in UTF8 „möglich“, wird die Datei als UTF8 eingestuft. Ist das nicht der Fall, wird sie als Latin1 (westeuropäische Encodierung) behandelt.

Dies Verfahren funktioniert sehr sicher. Es geht schief, wenn eine Datei nach Zusammenkopieren Abschnitte in unterschiedlichen Encodings enthält. GESStabs findet in einem solchen Fall Zeichen, die in UTF8-Encodierung unerlaubt sind. GESStabs entscheidet dann auf Latin1, sodass Zeichenketten in UTF8 sehr merkwürdig aussehen. In solchen Fällen muss der Benutzer, wie leicht einsehbar ist, selbst Hand anlegen. Das System kann auch nicht mit anderen ANSI Codepages als Westeuropa umgehen. Kyrillisch, Griechisch, Russisch und andere müssen zur Verarbeitung in UTF8 oder Unicode gewandelt werden.

Ebenso wie Script-Dateien können auch OPENQ-Dateien in verschiedenen Encodierungen gelesen werden. Das OPENQ-System beherrscht die Erkennung von Encodings[1].

Am einfachsten erscheint nach dem bisher Gesagten das Motto: Alles in UTF8. Leider geht es nicht so elegant weiter. Das vorherrschende Datenformat in der Sozial- und Marktforschung ist der spaltenfixierte Datensatz. Aus Performancegründen holt sich GESStabs einen Wert in einem Datensatz aus z.B. der 507. Spalte, in dem es ein Array adressiert. Wenn sich nun auf den Spaltenpositionen 1 bis 506 Alpha-Daten befinden, deren Buchstaben in UTF8 mehr als ein Byte belegen, geht das schief. Spaltenfixierte ASCII-Datensätze erwartet GESStabs immer in Latin1[2]. „Exotische“ Zeichen müssen entweder aus dem Script oder einem OPENQ-File stammen.

[1] Hier wird allerdings nur eine Teilmenge unterstützt: UTF8, UTF8BOM, UTF16LE, Latin1. Da die Antworten auf offene Fragen auch als Verbatim genutzt werden, und Daten oft von verschiedenen Feldinstituten stammen, ist dies manchmal fehlerträchtig. GESStabs unterstützt deshalb einen Modus, in dem man UTF8 fest einstellt, und GESStabs diese Encodierung während der Verarbeitung überwacht, siehe ENFORCEUTF8INOPENQ.

[2] Natürlich ist das unschön. UTF8-Texte als Bestandteil der Daten lassen sich wesentlich einfacher mit nicht-spaltenfixierten Formen verarbeiten. Hierfür bieten sich CSV-Dateien an, für die GESStabs auch UTF8 unterstützt. (siehe CSVINFILE)