Kurz erklärt: Verwechslungsgefahr bei Unicode-Zeichen

Eingefangen

Sven Krohlas

Wer Kunde von meinebank.de ist, klickt natürlich auf meinebank.de – und wird so vielleicht zum Opfer gewiefter Betrüger. Denn es könnte sich um zwei verschiedene Domainnamen handeln.

Betrüger versuchen auf immer wieder neue Art und Weise, Anwender auszutricksen. Derzeit hoch im Kurs: „Confusables" – Unicode-Zeichen, die anderen Zeichen und Symbolen ähneln oder gleichen und daher leicht verwechselt werden können (Homomorphien). So könnte die erste der beiden eingangs genannten Domains das kyrillische Zeichen „а" enthalten, das üblicherweise pixelgenau gleich dargestellt wird wie das lateinische Pendant „a". Solche verwechselbaren Zeichen können Angreifer in internationalisierten Domainnamen sowie in Mailadressen einsetzen – weitere Anwendungsfälle dürften folgen, insbesondere auch für die zahllosen grafischen Symbole (Emojis und Co.) in Unicode: Mit deren Zahl wächst die Gefahr, dass sich eines mit hinreichender Ähnlichkeit zu Bedienelementen einer Anwendung findet. Schleust ein Angreifer ein solches Zeichen ein – etwa einen grünen Haken – und die Anwendung zeigt nicht eindeutig an, welche Elemente der Anwender eingegeben hat und welche sie selbst nutzt, kann dies Verwirrung stiften.

Erhöhtes Risiko durch Internationalisierung

Einen Eindruck vom Potenzial der Confusables vermittelt ein Tool des Unicode-Konsortiums (siehe ix.de/z9w9). Es errechnet für den Namen des Autors dieses Artikels mehr als 10^{17} einander zum Verwechseln ähnliche Varianten.

Mehrere Gremien haben sich mit den Angriffsszenarien befasst, die die Vielfalt von Unicode mit sich bringt. Die Universal Acceptance Steering Group betrachtet die Anwendungsfälle Domains und internationalisierte Mailadressen, der Unicode Technical Report TR39 (Unicode Security Mechanisms) des Unicode-Konsortiums die allgemeineren Grundlagen. Der Report unterscheidet zwischen dem Täuschen menschlicher Leser und Ähnlichkeiten, die technische Schutzsysteme wie Spamfilter aushebeln sollen.

Bis auf die Knochen

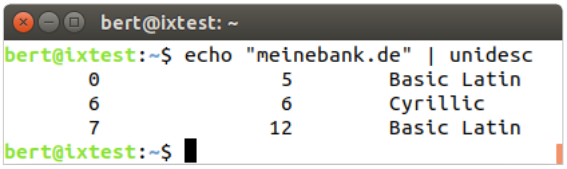
Mensch oder Maschine lassen sich auf zahlreiche Arten und Weisen in die Irre führen, zum Beispiel durch nicht angezeigte Zeichen. Allein davon kennt Unicode etliche, etwa die Zeichen zum Ändern der Schreibrichtung. Sie lassen sich sogar zu mehreren hintereinander anordnen, ohne das Aussehen des Textes zu verändern, und eignen sich daher wunderbar dafür, Anwender zu täuschen. Spamfilter, die Worthäufigkeiten auswerten (Bayes-Filter), stehen vor der Herausforderung, solche Symbole zu berücksichtigen – andernfalls sehen sie jede der unzähligen Unicode-Varianten eines Wortes als einmalig an.

Zum Erkennen ähnlicher Zeichen hat das Unicode-Konsortium Zeichen mit Homomorphien in Klassen eingeteilt und aus jeder Ähnlichkeitsklasse einen Vertreter zum sogenannten Prototyp ernannt. Auch Zeichenkombinationen können zu Ähnlichkeiten führen, zum Beispiel zwischen „rn" und „m". Auf Basis dieser Liste lassen sich sogenannte „Skelette" bilden: In dieser normalisierten Darstellungsform sind Zeichenketten identisch, deren Ursprungsformen einander nur ähneln. Der Standard nennt diese Strings „äquivalent". Da außerdem gleichbedeutende Unicode-Zeichen intern verschieden dargestellt sein können, gilt es, den String zunächst in eine Normalform (NFD) zu transformieren. Erst danach werden alle Zeichen durch ihren Prototyp ersetzt. Falls es zu einem Zeichen keine ähnliche Variante gibt, bildet es seine eigene Klasse und damit seinen eigenen Prototyp. Schließlich wird der String wieder NFD-normiert. Dieses Verfahren erkennt zumindest Confusables, die Menschen verwirren sollen.

Um es auch Filtersystemen zu ermöglichen, verschiedene Schreibweisen von V́íág̀rà zu erkennen, sind weitere Schritte notwendig, etwa diakritische Zeichen wie Akzente zu entfernen. Doch perfekt ist das Ergebnis nie, denn das menschliche Gehirn kann nach zahlreichen Wortveränderungen noch die beabsichtigte Bedeutung erkennen. Spam-Sammlungen zeigen unzählige Beispiele dafür.

Erst denken, dann klicken

Einen anderen Ansatz, Verwirrungsversuche zu detektieren, bietet der Blick auf den Wechsel von Schriften in einem String. Naturwissenschaftler nutzen laufend griechische neben lateinischen Zeichen in Formeln – in einem Domainnamen wäre dies verdächtig. Die IANA pflegt daher ein Verzeichnis von Praktiken bei der Verwendung internationalisierter Domainnamen, die im Rahmen der Registrierung zu prüfen sind. Im Zweifel sollte man eine Zeichenkette vor einem Klick untersuchen, beispielsweise mit den Unicode Utilities von Bill Poser (siehe Abbildung). Einen wirksamen Schutz vor Confusables bietet aber letztlich nur der gute alte Tipp, die eigenen Bookmarks zu nutzen oder die gewünschte Zeichenkette selbst einzutippen. (un@ix.de)



Ein Nebeneinander lateinischer und kyrillischer Zeichen ist höchst verdächtig.

Quellen

Tools und Standards zu Unicode Confusables finden sich unter ix.de/z9w9.

Sven Krohlas

ist E-Mail-Spezialist und IT Security Consultant bei der BFK edv-consulting GmbH in Karlsruhe.