ABBYY®

FALLSTUDIE

ABBYY® Historic OCR

Projekt Gutenberg-DE

Wie OCR-Software das kulturelle Erbe bewahrt

Für das Projekt Gutenberg vereinfacht ABBYY die Digitalisierung von Texten in Frakturschrift

Der Hintergrund

Vor über 15 Jahren initiierte die Internetagentur Hille & Partner aus Hamburg den deutschen Zweig des Projekts Gutenberg, um Werke der deutschsprachigen Literatur sowie klassische Übersetzungen fremdsprachiger Werke ins Deutsche online verfügbar zu machen. Zurzeit lassen sich über 5.500 Texte – Romane, Novellen, Theaterstücke, Gedichte sowie Sachbuchtitel – unter http://gutenberg.spiegel.de kostenlos abrufen. Henriette Davidis' Praktisches Kochbuch für die gewöhnliche und feinere Küche befindet sich ebenso darunter wie die Märchen der Marie de France, eine Beschreibung der Künstlerkolonie Worpswede von Rainer Maria Rilke oder Die Geschichte eines Braminen von Karoline von Günderode. In dieses Projekt nehmen die Mitarbeiter ausschließlich Autoren auf, die bereits 70 oder mehr Jahre verstorben sind, weil das Urheberrecht auf ihre Werke erloschen ist. Pro Monat wächst der Bestand um ungefähr 50 neue Bücher. Hille & Partner finanziert das Projekt unter anderem durch den Verkauf von Hörbüchern, CD-ROMs und DVDs aus dem Gutenberg-Bestand.

Die Herausforderung

Hille & Partner startete zunächst als Agentur, die klassisches Webdesign für Hamburger Verlage anbot. Das Projekt Gutenberg entstand quasi nebenher aus Liebhaberei, denn Hella Reuters, Inhaberin der Agentur, hegt eine große Leidenschaft für die deutsche Sprache und ihre Literatur. Sie wollte vor allem schöne Literatur einem größeren Publikum zugänglich machen: „Gerade im 18. Jahrhundert gibt es so viel zu entdecken, und die Freude an den Texten, die wollte ich mit anderen teilen." Zunächst tippten die Mitarbeiter von Hille & Partner einige seltenere Kinder- und Hausmärchen der Gebrüder Grimm ab und stellten die Texte ins Netz. Bis die Texte digital verfügbar waren, dauerte es ungefähr vier Wochen. Dass die Agentur das automatisierte Digitalisieren von Literatur überhaupt in ihr Portfolio aufnahm, ist einem glücklichen Zufall zu verdanken: Den Ausschlag dazu gab Ende der 90er Jahre der Büchergrossist Libri. Er beauftragte die Agentur, zahlreiche Klappentexte und Leseproben auf dem Computer verfügbar zu machten. Mit einfachem Abtippen war es nicht getan. Schon damals verwendete das Unternehmen Scanner und eine Spezialsoftware zum Erkennen von Buchstaben und Wörtern.

Über das Projekt Gutenberg

Das Projekt Gutenberg ist ein Online-Angebot, das die Internetagentur Hille & Partner ins Leben gerufen hat. Es bietet kostenlos Werke der deutschen Literatur im Internet an. Bisher sind über 5.500 Romane, Theaterstücke, Gedichte, Novellen und Sachbücher in deutscher Sprache von rund 1.200 Autoren ins Netz gestellt. Im Projekt Gutenberg werden ausschließlich Werke publiziert, deren Copyright bereits erloschen ist. Das Hosting hat SPIEGEL ONLINE übernommen. Neben den Mitarbeitern von Hille & Partner unterstützt eine Community von über 4.150 aktiven freiwilligen Helfern das Projekt.

Kontakt

Hille & Partner
Projekt Gutenberg-DE
Wandalenweg 5
20097 Hamburg
Deutschland
Tel.: 00 49 (0)40 - 899 75 55
reuters@ abc.de
http:// gutenberg.spiegel.de

ABBYY Europe GmbH
Elsenheimerstr. 49, 80687 München, Deutschland
Tel: +49 89 51 11 59 0, Fax: +49 89 51 11 59 59
sales_eu@abbyy.com
www.ABBYY.com
www.ABBYY.de

ABBYY® Historic OCR

Nachdem sich die Agentur in diesem Geschäftsfeld bereits einen Namen gemacht hatte, trat ein Verlag auf die Agentur zu und beauftragte sie mit dem Einscannen von Gerichtsdokumenten. Die Herausforderung bestand dabei, dass zahlreiche Texte in Fraktur gedruckt waren. Frakturschrift war seit Mitte des 16. bis Anfang des 20. Jahrhunderts die meistbenutzte Druckschrift im deutschsprachigen Raum - für Ende der 90er Jahre verfügbare OCR-Programme ein Ding der Unmöglichkeit, diese zu erkennen. Im Gegensatz zu den heute verwendeten Schriften gibt es oft keine einheitlichen Fonts für Frakturzeichen, was eine maschinelle Erkennung der Texte stark erschwert. Jedes einzelne Frakturzeichen existiert in vielen Varianten. Dazu kommen noch verschiedene Ligaturen und Schmuckzeichen, die das Druckbild von Dokument zu Dokument abweichen lassen. Besonders bei älteren Büchern beeinflussen darüber hinaus Schmutz, Stockflecken, Staubkörner oder Flusen die Qualität der Scans und damit das Erkennen der einzelnen Buchstaben und Wörter. Hella Reuters: „Für uns hieß das, dass wir die Scans stets nachbearbeiten mussten. Und das bedeutete jedes Mal einen Riesenaufwand."

Die Lösung

Damit ein Buch überhaupt digitalisiert werden kann, muss es Seite für Seite eingescannt werden. Als nächster Schritt werden die Scans mithilfe einer OCR-Software erkannt. Am Ende dieses Prozesses steht eine Datei in den gängigen Formaten, sei es Text only, Word, PDF, HTML oder XML.

Bis zum Jahr 2004 hatten Hella Reuters und ihre Kollegen verschiedene Produkte ausprobiert: „Aber die haben uns alle nicht zufriedengestellt. Doch dann hat uns ein Kunde auf ABBYY aufmerksam gemacht, und das war wirklich eine Offenbarung. Gerade was das Erkennen von Frakturschrift betrifft, schlägt ABBYY-Software seine Wettbewerber um Längen." Zunächst setzte Hille & Partner ABBYY FineReader XIX ein, später verwendete die Agentur ABBYY Recognition Server. Letzterer hat einen größeren Funktionsumfang und basiert zudem auf einer deutlich weiter entwickelten Technologie. Um Fehler weitgehend zu eliminieren, kombiniert die Software zwei Verfahren: die optische Textkennung und die Analyse auf Sprachebene.

Um die Standardzeichen zu erkennen, nutzt die OCR-Technologie verschiedene Muster und „Classifier", die in der Software hinterlegt sind. Hat die Software die einzelnen Zeichen erkannt, setzt sie sie zu ganzen Wörtern zusammen.

Nun beginnt die zweite Ebene der Analyse, denn neben der optischen Texterkennung werden auch Sprachmuster und Rechtschreibung einbezogen. Aber wie fast überall steckt der Teufel im Detail: In Deutschland wurde eine einheitliche Orthografie erst ab 1901 eingeführt. Bis dahin schrieb jeder Autor, wie er es für richtig hielt. Außerdem flossen zahlreiche Regionalismen in literarische Texte ein, einige Werke sind sogar auf Niederdeutsch verfasst. Das macht es schwer, auf eine standardisierte Rechtschreibung zurückzugreifen. Die Kombination aus zwei Analyseebenen sorgt dafür, dass die gescannten Texte nach der Bearbeitung kaum noch Mängel enthalten. Hella Reuters: „Ich bin wirklich erstaunt, wie exakt die ABBYY-Software arbeitet."

Obwohl die Fehlerquote sich über die Jahre verbessert hat, müssen auch die übriggebliebenen Erkennungsfehler entfernt werden. Diese Aufgabe übernimmt eine Community aus freiwilligen Helfern, die sich über das Internet-Portal GaGa (www. gaga.net) koordiniert. „GaGa" steht für „Gemeinsam an Gutenberg arbeiten". Hier lesen 4.150 aktive Helfer unentgeltlich die einzelnen Seiten Korrektur.

Das Ergebnis

Über 1,5 Millionen Seiten wurden seit Projektbeginn auf diese Weise digitalisiert und ins Netz gestellt. Jeden Monat kommen 50 neue Bücher hinzu. Dieses fantastische Ergebnis wäre ohne den Einsatz von ABBYY-Software nicht möglich gewesen. Insbesondere das nahezu fehlerfreie Erkennen von Frakturschrift hat die Bearbeitung von älteren Büchern für das Projekt Gutenberg enorm erleichtert und dadurch beschleunigt. Da die Werke nun digitalisiert sind, stehen sie online aber auch für elektronische Ausgabemedien, wie beispielsweise Amazon Kindle™ oder das Apple® iPad® zur Verfügung, was den Nutzerkreis deutlich erweitert.

Zu Beginn des Projektes hätte sich Hella Reuters nicht vorstellen können, dass es einmal eine derartige Resonanz haben würde. Aber wenn die Hamburgerin auf ABBYY zu sprechen kommt, kann sie ihre Begeisterung kaum zurückhalten: „Es ist wirklich gewaltig, was die Software von ABBYY für das Projekt Gutenberg geleistet hat. Nein, das ist keine Lobhudelei, es ist wirklich grandios, ohne ABBYY gäbe es das Projekt Gutenberg nicht."

Über ABBYY

ABBYY ist ein führendes Unternehmen in der Entwicklung von Technologien für Dokumenterkennung, Dokumentumwandlung, Data Capture und Linguistik.

Die Abteilung Forschung & Entwicklung widmet sich der kontinuierlichen Weiterentwicklung und Verbesserung der OCR-Technologien, um eine noch genauere Erkennung einer immer breiteren Masse von Dokumenten zu erreichen. Seit dem Jahr 2003 ist ABBYY in die Entwicklung von Fraktur OCR involviert, einer speziellen Technologie für die Erkennung von gebrochenen Schriften. Kunden und Partner profitieren von ABBYYs Forschungsarbeit durch die Verfügbarkeit der Technologieentwicklungen in den neuesten Produktversionen.

Zum Produktportfolio von ABBYY gehören: ***FineReader OCR*** *und* ***PDF Transformer*** *– Endanwenderprogramme zur Umwandlung von Dokumenten;* ***Recognition Server*** *– eine serverbasierte Lösung für OCR und PDF-Umwandlung;* ***FlexiCapture*** *– Data Capture Lösung zur Verarbeitung von Formularen, semi- und unstrukturierten Dokumenten;* ***FineReader Engine SDKs*** *mit dem gesamten Leistungsumfang der ABBYY OCR-Technologien;* ***Lingvo*** *– eine Serie von elektronischen Wörterbüchern.*

Mehr Informationen über ABBYY unter www.ABBYY.com

ABBYY®

ABBYY Europe GmbH
Elsenheimerstr. 49, 80687 München, Deutschland
Tel: +49 89 51 11 59 0, Fax: +49 89 51 11 59 59
sales_eu@abbyy.com
www.ABBYY.com
www.ABBYY.de