bitfarm® ARCHIV

Dokumentenmanagement

OCR-Leistung im Vergleich Tesseract - Omnipage

Inhaltsverzeichnis

I. Texterkennung im elektronischen Dokumentenmanagement

1. Wozu Texterkennung?

Texterkennung findet im elektronischen Dokumentenmanagementsystem (DMS) dort Anwendung, wo papiergebundene Dokumente digitalisiert und deren Inhalte vom Computer auswertbar und durchsuchbar gemacht werden sollen. Nach dem Scanvorgang liegt ein Dokument als digitale Rastergrafik vor. In der Darstellung auf den üblichen Anzeigegeräten entspricht es weitestgehend dem Original. Die Verarbeitung von Informationen findet im Computer allerdings in einem codierten Zeichen- und Nummernsatz statt (bspw. ASCII oder Unicode). Texterkennungssoftware (OCR) ist in der Lage, aus Rastergrafiken die korrespondierenden maschinencodierten Informationen zu extrahieren.

Im Dokumentenmanagementsystem werden diese Informationen dann in eine Datenbank geschrieben. Die Suchfunktionen der Datenbank gewährleisten den schnellen Zugriff auf Dokumente über beinhaltete Text- oder Zahlenbestandteile zu jedem Zeitpunkt mit hoher Geschwindigkeit. Des Weiteren können die extrahierten codierten Informationen zur weiteren Klassifizierung und Verarbeitung verwendet werden. Als typisches Beispiel ist das Einscannen einer Rechnung mit basierend auf dem OCR-Ergebnis automatischer Erkennung des richtigen Ablageortes im DMS, Benachrichtigung des zuständigen Sachbearbeiters und Extraktion der Kopfdaten wie Rechnungsnummer und Betrag zu nennen, die auch nachgeschalteten Softwaresystemen (z. B. FiBu) zur Verfügung gestellt werden.

2. Probleme bei der Texterkennung

Maschinelle Texterkennung ist ein aufwendiger Prozess. Gleichzeitig ist er fehleranfällig. Für die Erkennung eines Buchstabens oder einer Zahl aus einer Rastergrafik wird daher innerhalb der OCR eine Wahrscheinlichkeit angegeben, die nie 100% beträgt. Am Beispiel der Unterscheidung zwischen den Zeichen O und 0 wird dies schon für einen Menschen zum Problem. Hier greifen wir, und auch die OCR auf den umgebenden Kontext zurück, um zwischen O (wie in „Oh“) und 0 („Null“) unterscheiden zu können. Was ist aber mit einer Zeichenfolge ohne hinreichenden Kontext, z. B. „rtHs043R25“? Verbessern ließe sich dies durch eine geeignete Schriftart. Die Courier-Gruppe beispielsweise beinhaltet die entsprechenden Merkmale, weshalb sie oft für Beschriftungen genutzt wird, die sowohl der Mensch, als auch der Computer fehlerfrei erkennen sollen. Beispiel: KFZ-Kennzeichen.

SI-BF 123

Die Wahrscheinlichkeit für Fehlerkennungen steigt naturgemäß zusätzlich mit schlecht leserlichen Vorlagen. Alte, vergilbte und verwaschene Dokumente, die auch ein Mensch nicht mehr entziffern kann, werden vom Computer nach dem aktuellen Stand der Technik auch nicht erkannt werden, geschweige denn fehlerfrei. Ungeeignete Scangeräte (im Ergebnis beispielsweise „schiefe“ Rastergrafiken, schlechte Schwellwerte und damit störendes Rauschen) tragen ebenfalls einen großen Teil zu Fehlern in der Texterkennung bei – ebenso wie die Qualität der verwendeten Texterkennungssoftware.

Gleichwohl wird Texterkennung immer fehlerbehaftet bleiben, weshalb relevante Entscheidungsprozesse nie allein auf Basis eines Texterkennungsergebnisses gefällt werden dürfen. Die GoBD schreibt daher beispielsweise die Sichtprüfung und den Vergleich bei relevanten Ergebnissen der Texterkennung zwingend vor. Der „Xerox-Bug“ von 2013 dokumentiert eindrucksvoll und erschreckend zugleich die Konsequenzen von „blindem Vertrauen“ in die Ergebnisse von Texterkennungssoftware.

3. Texterkennung in bitfarm-Archiv DMS

Texterkennung ist ein wichtiger Bestandteil von bitfarm-Archiv in allen Versionen. Die Verfügbarkeit passender Scangeräte und OCR-Software, sowie schneller Datenbankprogrammierung, die auf aktueller Hardware Suchzeiten von weniger als einer Sekunde in hunderttausenden Dokumenten ermöglichen, sind ein Markenzeichen des bitfarm-Archiv Dokumentenmanagementsystems. Die Ablage von Dokumenten kann dabei denkbar einfach erfolgen und eine mühevolle händische Verschlagwortung ist nicht notwendig, um das Dokument trotzdem später schnell wiederzufinden. Einzelne OCR-Fehler sind hier verschmerzbar, da bei einer Suche dem Anwender der gesamte Text in Kombinatorik und damit viele verschiedene Wege zur Verfügung stehen, ein Dokument zu erreichen. Funktioniert ein bestimmter Suchstring nicht, gibt es einen anderen, der das Dokument liefert. Das funktioniert natürlich nur, wenn ausreichend Text / Zahlen auf dem Dokument vorhanden sind und diese auch mit einer gewissen Genauigkeit erkannt wurden. bitfarm-Archiv bietet ein internes Kontrollsystem, welches die Qualität der Dokumente misst und im Bedarfsfall händische Verschlagwortung anfordert.

Dieses Konzept gestattet es, dass OCR-Fehler im Einzelfall auftreten dürfen, ohne zu größeren Problemen zu führen. Der Wegfall der „Kosten“ händischer Verschlagwortung gerade bei großen Dokumentenmengen rechtfertigt den möglicherweise auftretenden geringen Mehraufwand bei einzelnen Suchen.

Dennoch sollte die Texterkennung eine möglichst hohe Qualität aufweisen, damit direktes Suchen und Finden in nahezu allen Fällen möglich ist.

Während bei der volltextbasierenden Suche von Dokumenten einzelne Lesefehler tolerabel sind, ist dies bei der automatischen Extraktion von Metadaten (automatische Verschlagwortung) korrekturbedürftig. Metadaten, wie zum Beispiel Rechnungsnummer, Auftragsnummer, Zahlbetrag im Falle von kaufmännischen Dokumenten, werden häufig als führende Merkmale für den Beleg verwendet und in anderen Softwaresystemen weiterverarbeitet. OCR-Lesefehler bei diesen Metadaten müssen deshalb händisch korrigiert werden, was mit entsprechendem Aufwand verbunden ist.

bitfarm-Archiv setzt in der aktuellen, wie auch in allen vorherigen Versionen der Enterprise-Edition die Omnipage-OCR des Herstellers Kofax (vormals Nuance und Scansoft) ein. Verschiedene Tests in der Fachliteratur (c't 22/2005, c't 16/2018) bescheinigen der Software im Vergleich mit anderen Produkten beste Ergebnisse nach dem Stand der Technik.

Währenddessen zeigt das Projekt „Tesseract“ eine interessante Entwicklung. Tesseract ist eine Texterkennungssoftware, welche ursprünglich bei Hewlett-Packard entwickelt wurde und in der Folgezeit ihren Weg zu Google und unter die Apache Lizenz (einen Open-Source Lizenztyp) gefunden hat. Tesseract setzt in seinen neuesten Versionen so genannte „neuronale Netze“ zur Verbesserung der Erkennungsleistung ein und erzielt damit beachtliche Erfolge.

Die GPL-Version von bitfarm-Archiv wird seit 2020 mit Tesseract ausgeliefert und auch für die Enterprise-Version gibt es die Möglichkeit, Tesseract als OCR-Engine zu verwenden. Wir wollen daher beide Texterkennungslösungen hier einmal gegenüberstellen. Dabei interessiert uns zum einen die Qualität des extrahierten Textes bei unterschiedlichen Dokumentenarten und –Qualitäten, zum anderen eine Geschwindigkeits- und Kostenbetrachtung. Abschließend wollen wir auf Basis unserer Ergebnisse eine fundierte Empfehlung für verschiedene Anwendergruppen aussprechen.

II. Ein Vergleich Tesseract – Omnipage OCR der bitfarm Volltexterkennung

Erste Pilotprojekte im Enterprise-Bereich mit der Tesseract OCR sorgten mit guten Ergebnissen zur Zufriedenheit der Anwender. Tesseract ist im Gegensatz zur Omnipage OCR Open-Source, was für bitfarm-Archiv als ebenfalls Open-Source DMS erweiterte Nutzungsmöglichkeiten bietet. In Planung ist beispielsweise eine verteilte OCR über die zur Verfügung stehende, nicht genutzte Rechenleistung der PCs im lokalen Netz des Kunden, um täglich sehr große Dokumentenmengen verarbeiten zu können.

Hinzu kommt, dass Tesseract sehr aktiv weiterentwickelt wird. Bei Omnipage hingegen konnte in den letzten Jahren keine Steigerung der OCR-Leistung mehr festgestellt werden – hier hat man den zugegeben schon sehr guten Stand konserviert und lediglich an der Oberfläche Modellpflege betrieben.

Unsere Motivation bei bitfarm-Archiv ist, jederzeit dem Kunden die aus technischer und wirtschaftlicher Sicht optimale Lösung bereitzustellen. Tesseract scheint vielversprechend, aber wie gut ist es wirklich im Vergleich und welche Empfehlung gilt für welchen Anwendungsfall? In diesem Test stellen wir den beiden OCR-Lösungen verschiedene praxisnahe Aufgaben und vergleichen die Ergebnisse.

1. Versuchsaufbau

Wir verwenden Testdokumente aus einem Satz uns zur Verfügung stehender Realdokumente. Es handelt sich um insgesamt 60 Rechnungen von ca. 40 verschiedenen Kreditoren. Somit liegt eine umfangreiche Dokumentensammlung mit einer großen Anzahl verschiedener Layouts und Schriftarten vor. Alle Dokumente wurden eingescannt und sind von unterschiedlicher Qualität, d. h. per Sichtprüfung erkennt man deutliche Unterschiede der Dokumentenqualität mit einem Bereich von sehr gut und nahezu makellosen Scans bis hin zu schief eingescannten Dokumenten, die unvollständige Stempel und auch nicht auf dem Originaldokument enthaltene Pixel (z. B. durch verunreinigte Scanner) enthalten.

Um eine möglichst objektive Testreihe durchzuführen, wurden jeweils im Wechsel Stapel von Dokumenten mit Tesseract und Omnipage in ein bitfarm-Archiv Testsystem archiviert und dann verschlagwortet.

2. Verschlagwortung

Zur Erstellung der Verschlagwortungsregeln wurde sowohl die Methode mittels *searchdirection*, *getstrings* als auch reguläre Ausdrücke (*searchpattern*) verwendet. Dabei wurden im Wechsel TESS/OMP einmal komplett neue Regel erstellt, als auch bereits erstellte Regeln für die andere OCR-Engine benutzt.

```
1623 naming section OMP Rechnungsnummer Berufsbildungsze
1624 archivtabelle=OCR_Test
1625 andindocument=Berufsbildungszentrum (bbz)
1626 searchstring=Rechnungsnummer:
1627 searchsteps=1
1628 getstrings=1
1629 nodeletedashes
1630 zusatzfeld=Rechnungsnummer
1631 end section
```

Beispiel einer Regel für die automatische Verschlagwortung in bitfarm-archiv

3. Bewertung

Die Erstellung Verschlagwortung wurde mit leicht, mittel und schwer bewertet.

Leicht: a) Regel konnte auf Anhieb gefunden werden oder b) Regelübernahme von OMP nach TESS mittels keiner oder nur minimaler Korrektur möglich

Mittel: Es brauchte mehrere Durchläufe, um den gewünschten Wert zu finden bzw. korrekt auszulesen

Schwer: Nach mehreren Versuchen musste ein *Searchpattern* geschrieben werden, da nur ein regulärer Ausdruck den Wert (korrekt) auslesen kann

Nicht möglich: Der Wert kann nicht erfasst werden. Im Test nur 1-mal aufgetreten

Fehleranzahl: die Anzahl der nicht, oder nicht korrekt erkannten Werte pro Dokument. Gesucht wurden folgende Werte, um damit Zusatzfelder automatisch zu befüllen: Rechnungsdatum, Rechnungsnummer, Kundennummer, Betrag.

	leicht	Fehler	mittel	Fehler	schwer	Fehler	nicht möglich	Fehler
TESS	41	5	15	11	3	0	1	1
OMP	41	2	10	3	9	9	0	0

Bewertung der Verschlagwortungsregeln

4. Beurteilung

Zunächst fällt auf das Omnipage und Tesseract an unterschiedlichen Dokumenten bzw. Textstellen an Ihre Grenzen kamen. Es traten mehrfach OCR Erkennungsfehler bei der einen Software auf, während dieselbe Textstelle von der anderen OCR Lösung jedoch fehlerfrei erkannt wurde.

Tesseract erstellt Volltext von gescannten Rechnungen mit deutlich mehr Zeilenumbrüchen. Der gesuchte Wert ist also vorhanden und kann somit ausgelesen werden, es ist jedoch möglich, dass das Finden der Verschlagwortungsregel aufwendiger als mit Omnipage ist, da z. B. der Begriff „Rechnungsnummer“ und die im Originaldokument dahinterstehende Rechnungsnummer, also der gesuchte Wert, nicht zusammenhängend bzw. hintereinander im Dokument zu finden sind. Es gab auch ein Dokument der Testreihe, wo der gesuchte Wert zwar im Volltext vorhanden war, jedoch aus dem Zusammenhang „verschoben“ wurde, sodass es nicht möglich war, den Wert mit einer Regel zuverlässig auszulesen. Doch dieser Fall war die Ausnahme.

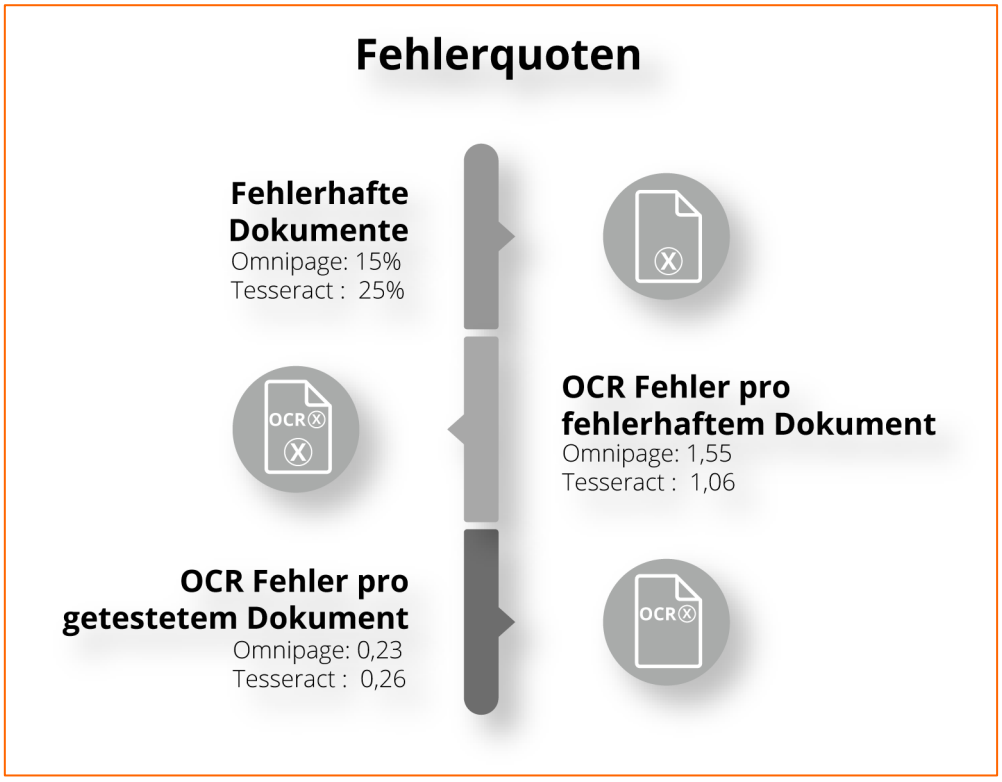
5. Fehlerquoten

a) Omnipage

- 9 Problem-Dokumente mit insgesamt 14 OCR Fehler
- 1,55 Fehler je fehlerhaftem Dokument
- *0,23 Fehler je getestetem Dokument, d. h. alle 4,3 Dokumente kam es zu einem Fehler*

b) Tesseract

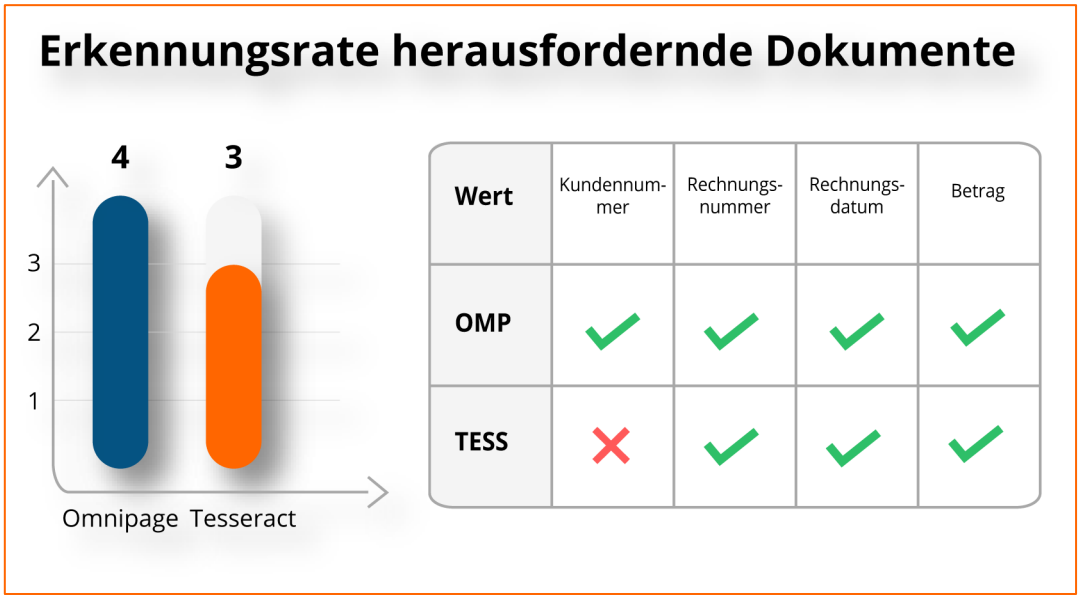
- 15 Problem-Dokumente mit insgesamt 16 OCR Fehler
- 1,06 Fehler je fehlerhaftem Dokument
- *0,26 Fehler je getestetem Dokument, d. h. alle 3,8 Dokumente kam es zu einem Fehler*



Allgemeine Fehlerquoten

6. Qualitativer Vergleich von herausfordernden Dokumenten

Für diesen Vergleich wurde ein schief eingescanntes Dokument mittlerer Qualität genommen und die OCR Ergebnisse verglichen. Im Test wurde dann nach 4 Standardwerten gesucht: Kundennummer, Rechnungsnummer, Rechnungsdatum und Betrag.



Erkennungsrate von Standardwerten bei herausfordernden Dokumenten

Bei der Suche nach den vier buchungsrelevanten Werten erkannte Tesseract ein Zeichen der Kundennummer falsch, während Omnipage alle vier korrekt erfasste. Zusätzlich wurde auch der komplette Volltext, also nicht nur die gesuchten Werte, miteinander verglichen. Tesseract erkannte stellenweise Zeichen, die Omnipage nicht erkennen konnte, und Omnipage war an anderen Stellen korrekter als Tesseract. Fairerweise muss man auch betrachten, wie wertvoll der erkannte Volltext letztendlich für die Suche im DMS ist, oder das Befüllen von Feldern. Z. B. wurde das Datum (03. Jan 2017) des Eingangsstempels von OMP nicht erkannt, das Ergebnis von TESS mit „3, Jan 2017“ ist jedoch nur teilweise irrelevant, da dieses Datum im Volltext zunächst nur von einem Menschen identifiziert werden kann und für eine Volltextsuche unbrauchbar ist. Das Befüllen eines Datumsfeldes mit diesem Wert wäre allerdings möglich, dabei wird der Wert in ein Standarddatumsformat umgewandelt und steht somit einer Datumssuche über ein Datumsfeld zur Verfügung.

In einem weiteren Vergleich wurde eine Ingram Rechnung mittlerer Qualität, mit deutlich sichtbaren Streifen und anderen Fehlern auf dem Dokument durch beide OCR Programme geschickt. Relevante Werte, die Fehler enthielten:

Fehleranzahl		Erfasster Wert		Typ
TESS	OMP	TESS	OMP	
0	1	PC667	2C667	Artikelbezeichnung
1	0	TSAGJEV30	TS4GJFV30	Hersteller Nr.
0	1	KVR33	-rzsp-33,	Artikelbezeichnung

An diesen Beispielen erkennt man, wie beide Produkte an unterschiedlichen Stellen Fehler bei der Texterkennung machen können. Komplett fehlerfrei erkannte Dokumente sind eher die Ausnahme in der Testreihe gewesen.

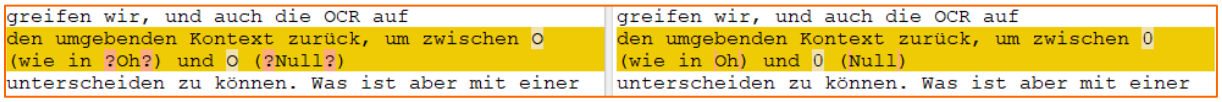
7. Qualitativer Vergleich von Dokumenten bester Qualität

Um den Vergleich der beiden OCR Lösungen abzuschließen, haben wir uns auch die Leistung beider Programme angeschaut, wenn Dokumente eingelesen werden, die digital und in bester Qualität erstellt wurden. Hierfür wurden verschiedene Dokumente mit dem bitfarm Archivdrucker gedruckt, dabei in TIFFs umgewandelt und dann durch die OCR geschickt. Die Dokumente weisen bei einer Sichtprüfung mit einer Vergrößerung von 200% ein perfektes Schriftbild ohne erkennbare Fehler auf.

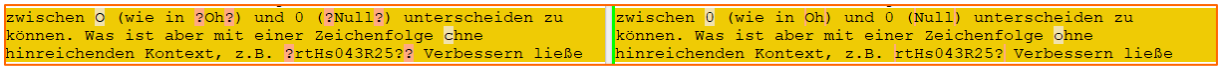
Bemerkung: Bei der Versuchsanordnung haben wir zunächst die Vorgehensweise aus Kapitel 6 wiederholt und gescannte Rechnungen archiviert. Da sich dabei herausstellte, dass es hier aufgrund der geringen Zeichenanzahl von Rechnungsdokumenten zu keinen zuverlässigen, aussagekräftigen Ergebnissen kommt, sind neue Testdokumente herangezogen worden, um Volltextdaten aus mehrseitigen Dokumenten mit mehr als 1000 Wörtern zu erhalten. Es wurde beim Testen ausschließlich auf die von der OCR Software nicht bzw. falsch erkannten Zeichen geachtet, und nicht auf Fehler bei der Verschlagwortung, die aufgrund mangelhafter OCR Daten auftraten.

Im ersten Lauf ist ein Auszug einer Publikation des Grundgesetzes verwendet worden. Hier wurden von beiden Programmen alle Wörter fehlerfrei erkannt. Lediglich kleine Unterschiede bei Zeilenumbrüchen und Leerzeichen wurden in den Volltext geschrieben. Ein gefundener Unterschied war z. B. dass aus „2 a“ mit Tesseract „2a“ wurde. Da ansonsten keine relevanten Unterschiede auftraten und diese Textstelle im Kontext unbedeutend und ein Einzelfall war, wurde der Fehler als irrelevant eingestuft.

Bei einem weiteren Testdurchlauf haben wir diesen für den Vergleich erstellten Bericht mit ca. 2300 Wörtern und Schrift *Calibri Light* mit beiden OCR Programmen gescannt. Für einen zweiten OCR-Scan wurde die Schriftart in *Courier New* geändert und das Dokument erneut als PDF durch die OCR Erkennung geschickt.



Calibri Light: links Tesseract, rechts Omnipage



Courier New: links Tesseract, rechts Omnipage

Wie man hier sieht, haben beide Programme Probleme, an dieser speziellen Textpassage zuverlässig zwischen Buchstabe O und Ziffer 0 zu unterscheiden, und obwohl Tesseract mit der Courier Schriftart im ersten Satz korrekt differenziert und Omnipage nicht, tritt danach ein Fehler im nächsten Satz auf. Das Dokument enthält über 2300 Wörter und bei beiden Programmen sind je nach Schriftart keine Fehler bzw. 4 Fehler aufgetreten. 2 davon waren Fehler bei Groß- und Kleinschreibung und somit für eine Standardsuche im DMS irrelevant (Ergebnis wird trotzdem gefunden). Der Fehlerquotient ist hier 0,0017 Fehler pro Wort, oder anders ausgedrückt kommt es alle 575 Wörter zu einem Fehler.

Fehlerquoten Tesseract und Omnipage bei Dokumenten von bester Qualität

Fazit: OCR Software arbeitet nicht hundertprozentig genau, doch die Ergebnisse sind in der Praxis sehr brauchbar. Tesseract und Omnipage lieferten im Test gleichwertige Qualität.

8. Verarbeitungsgeschwindigkeit

Im Zuge des Vergleichs der beiden Softwarelösungen wurde auch die Geschwindigkeit der OCR Erkennung untersucht. Ein Testdokument mit 480 Seiten wurde auf verschiedenen Servern jeweils mit Tesseract und Omnipage archiviert. Die Verarbeitungszeiten der OCR wurden dabei von der bitfarm Archivierung protokolliert.

Omnipage kann maximal vier CPU-Kerne für die Verarbeitung mehrseitiger Dokumente benutzen und verwendet dabei einen CPU-Kern pro Seite. Bei Verwendung von Tesseract wird ebenfalls pro CPU-Kern eine Dokumentenseite verarbeitet, die Begrenzung auf vier Kerne entfällt dabei jedoch.

Es kann nur ein OCR Prozess, also nicht mehrere parallel, laufen, was bedeutet das alle Dokumente nur nacheinander verarbeitet werden. Die Verwendung mehrerer Kerne kommt daher ausschließlich bei mehrseitigen Dokumenten zum Einsatz.

Verfügt der Server über mehr als vier Kerne, kann Tesseract bei vielen mehrseitigen Dokumenten (>4 Seiten) ihre bessere Skalierbarkeit ausspielen und auch die Performance der Omnipage dadurch übertreffen. Für einen aussagekräftigen Vergleichstest durften wir somit nicht mehr als vier Kerne einsetzen, um die maximal möglichen CPU-Kerne von Omnipage nicht zu übersteigen.

In der Standardkonfiguration mit einem Server brauchte Tesseract im Test 1,7 bis 2-mal so lange wie Omnipage. Hierbei wurden mehrere Durchläufe gemessen und dabei Server mit zwei und vier Kernen eingesetzt.

Im Performancetest der noch in der Entwicklung befindlichen, verteilten Tesseract OCR wurde mit 15 beteiligten Clients (durchschnittliche Office Workstations) das Testdokument fast 10x schneller verarbeitet als mit Omnipage. Dieses Testergebnis wird als repräsentatives Beispiel aufgeführt. Die tatsächliche Performance ist abhängig von der Leistung der eingesetzten CPUs, und die Anzahl der zur Verfügung stehenden CPU-Kerne.

In einem weiteren Feldversuch der verteilten Tesseract wurden im Verbund von mehreren Workstations mit insgesamt 30 unterschiedlichen CPU-Kernen 500.000 Einzelseiten in 24 Stunden mit Tesseract OCR verarbeitet. Auch dieses Ergebnis muss als demonstratives Beispiel betrachtet werden, nicht als Referenzergebnis. Dafür müssten immer auch CPU-Typen/-Taktraten, Aufbau der Dokumente, Netzwerk-Performance etc. mit in Betracht gezogen werden.

Zwar können auch mit Omnipage mehrere Verarbeitungsserver (oder Multiqueue) eingesetzt werden, hier werden jedoch immer nur komplette Dokumente verteilt. Beim Einsatz von Tesseract werden alle einzelnen Seiten eines Dokumentes zur OCR verteilt, simultan verarbeitet und danach wieder zusammengesetzt.

Das heißt jedes mehrseitige Dokument wird vom Cluster verarbeitet und profitiert von der kollektiven Verarbeitungsgeschwindigkeit, die sich aus den freien Ressourcen der vorhandenen Client-PCs zusammensetzt. Serverinstallationen wie bei der verteilten OCR mit Omnipage sind also in Zukunft bei Tesseract für hohen OCR-Durchsatz nicht mehr zwingend notwendig.

Lizenzkosten entfallen für Tesseract vollständig.

Beim Aufbau eines Omnipage Multiqueue Verbunds dagegen muss jede Installation lizensiert werden.

9. Zusammenfassende Interpretation der Ergebnisse und Empfehlungen

Die Ergebnisse sind durchaus knapp, dennoch liefert Omnipage eine etwas bessere Erkennungsleistung. Bei schwierigen Dokumenten, deren Scanbild nicht wirklich optimal daherkommt, ist die Omnipage OCR etwas besser als Tesseract, d. h. liefert weniger Fehler. Dennoch: Die Omnipage-Engine wurde in den letzten 10 Jahren nicht mehr weiterentwickelt. Das macht zwar das gute Produkt nicht schlechter, aber die Tendenz spricht eine klare Sprache dafür, dass Tesseract alsbald die Führung übernehmen wird, wenn deren Entwicklung so weiter verläuft. Nahezu auf Augenhöhe operiert Tesseract jetzt schon.

Omnipage und Tesseract liegen in der Erkennungsqualität nicht sehr weit auseinander. Jedoch zeigen die ausgelesenen Volltexte signifikante Unterschiede im Hinblick auf das Layout (Zeilenumbrüche, Leerzeichen, Tabstops). WFD-Regeln nutzen jedoch häufig die spezifischen Layouts, um Metadaten aus Dokumenten auszulesen. Bestehende WFD-Regeln sind also möglicherweise nicht kompatibel zwischen einer Omnipage oder einer Tesseract basierten Texterkennung. Der Wechsel der Texterkennungskomponente ist im Falle von größeren WFD-Regelwerken für die automatische Verschlagwortung also mit einigen Anpassungsarbeiten verbunden. Deshalb muss ein Wechsel der OCR-Engine kritisch im Kosten-/Nutzenverhältnis geprüft werden.

Bei Neuinstallationen würden wir hingegen auf Grund der guten allgemeinen Texterkennungsqualität, der hohen Geschwindigkeit, die in Zukunft noch weiter gesteigert werden kann, und der variablen Anpassbarkeit an weitere Schriftarten durch immer mehr verfügbaren Trainingsdaten und nicht zuletzt durch das Fehlen von Lizenzkosten zur Tesseract-OCR raten.

Omnipage empfehlen wir weiterhin dort, wo es auf das technisch bestmögliche OCR-Ergebnis ankommt und sowohl Geschwindigkeit als auch Lizenzkosten eine untergeordnete Rolle spielen.

Gleichwohl empfehlen wir allen Kunden, zunächst für ein optimales Scanergebnis zu sorgen. Dokumentenscanner wie die von uns empfohlenen Modelle von Alaris oder Fujitsu sind die wichtigste Voraussetzung für eine zufriedenstellende Texterkennung.

Es liegen Welten zwischen den Ausgaben eines Multifunktionsgerätes mit Scanfunktion und eines spezialisierten Dokumentenscanners. Weitere Informationen zu diesem Thema finden Sie hier:

Blog Beitrag

Glossar

Bei Fragen oder Problemen, wenden Sie sich an den bitfarm-Softwaresupport
Telefon: +49 (271) 31396-0
E-Mail: support@bitfarm-archiv.de
Telefonische Erreichbarkeit des Supports:
Mo. – Do. 8:00 Uhr bis 17:00 Uhr
Fr. 8:00 bis 15:00 Uhr

Bitte halten Sie Ihre Vertrags- oder Partner-Nummer bereit.