Forensische Analyse von .docx-Dateien zur Plagiatserkennung

Philipp Hagemeister, philipp.hagemeister@uni-duesseldorf.de

Um von Ghostwritern erstellte Arbeiten erkennen zu könne, können mithilfe des *dplagiat*-Tools Metadaten aus docx-Dateien extrahiert werden. Wegen der Ähnlichkeiten der Dokumentenformate ist dplagiat weiterhin auch auf von docMicrosoft Word Binary-Dateien konvertierte docx-Dateien (mit leichten Informationsverlusten) anwendbar.

dplagiat ist bisher nur auf der Kommandozeile ausführbar und erstellt dann wahlweise eine HTML-Ausgabe-Datei oder fasst die Ergebnisse direkt zusammen. Der Quellcode ist unter https://github.com/phihag/dplagiat unter den Bedingungen der AGPL frei einseh-, verwend- und veränderbar.

Sicherheitsabschätzungen

Alle mit dplagiat getroffenen können nur unter einigen Annahmen getroffen werden, nämlich das die analysierten Dateien weder unbeabsichtlich noch beabsichtigt manipuliert wurden. Prinzipiell könnten z.B. Software-Fehler (relativ unwahrscheinlich) oder Bitfehler (sehr unwahrscheinlich) auftreten. Da die `rsidr`-Werte, auf die die Analyse aufbaut, zufällig ausgewürfelt werden, kann es zu Kollisionen kommen, aufgrund deren zwei Segmente als eines angezeigt werden würden. Die Wahrscheinlichkeit hiervon hängt von der Anzahl der Segmente ab und schwankt für die bisher untersuchten Dokumente zwischen 0,0001 und 0,1 Prozent. Zusammenfassend sind solche Zufallsfehler unwahrscheinlich.

Allerdings ist es ohne weiteres möglich, docx-Dokumente beabsichtigt zu manipulieren und die Metadaten zu fälschen. Dafür sind noch nicht einmal spezielle Programmierkenntnisse oder Programme erforderlich; schon mit auf praktisch jedem Betriebssystem vorhandenen Editoren und Tools können alle untersuchten Informationen gefälscht werden. Dies kann leider nur sehr beschränkt erkannt werden. Solche Unregelmäßigkeiten wurden nicht gefunden.

Ausgabe

Die wichtigsten von *dplagiat* erfassten Daten sind die Anzahl der Speichervorgänge, die `rIdr`-Metadaten, sowie das Erstellungs- und das Bearbeitungsdatum. Die nachfolgende Tabelle zeigt, welche Auswirkungen häufige Operationen auf diese Daten (mit Microsoft Office) haben:

Aktion	#Speichervorgänge	`rIdr`	Erstellungsdatum	Bearbeitungsdatum
Erstellen einer Datei	=1	-	Wird gesetzt	Wird gesetzt
Hinzufügen/ Ändern von Text	-	rIdr-Zuweisungen	-	-
Speichern	+1	neue rIdrs	-	Wird gesetzt
Speichern unter neuem Namen	=2	-	Wird gesetzt	Wird gesetzt
Einfügen von kopiertem Inhalt	-	Neuer Inhalt taucht unter **einer** rIdr auf	-	-
Umbenennen im Explorer	-	-	-	-

Bewertung

Mit diesen Informationen kann also die Bearbeitungshistorie approximiert werden. Bei einem normalem Dokument in Bearbeitung würde man ein frühes Erstellungsdatum, eine große Anzahl von Speichervorgängen sowie viele durch `rIdrs` bezeichnete Sektionen von variabler, aber kleiner Größe erwarten. Wenn der Student kurz vor Abgabe die Datei unter einem neuem Namen abgespeichert hat, sind späte Erstellungs-/Bearbeitungsdaten und eine vergleichsweise niedrige Zahl von Speichervorgängen zu erwarten.

Wenn allerdings das Dokument durch eine oder zwei gigantische Sektionen dominiert wird und nur selten gespeichert wurde, dann muss der Inhalt aus einer anderen Datei kopiert worden sein. In der Abgabehektik ist es sicherlich möglich, dass Studenten das gesamte Dokument oder große Teile in eine neue Datei kopieren und dann abspeichern. Wenn allerdings die Datei schon vor Monaten erstellt wurde, der gesamte Inhalt aber aus einer anderen Datei kam, ist die einzig naheliegende Interpretation, dass die Datei auf Vorrat erstellt wurde und die eigentliche Arbeit in einer anderen Datei statt fand. Diese Angabe wäre jedoch jedoch mindestens erklärungsbedürftig; und deutet sonst auf das Verfassen des Dokuments auf einem anderem Rechner, also wahrscheinlich von einer anderen Person hin.