Forensische Analyse von .docx-Dateien zur Plagiatserkennung

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

Um von Ghostwritern erstellte Arbeiten zu erkennen, können mithilfe des *dplagiat*-Tools Metadaten aus docx[1]-Dateien extrahiert werden. Wegen der Ähnlichkeiten der Dokumentenformate ist dplagiat weiterhin auch auf von doc[2]-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.

Ausgabe

Die wichtigsten von *dplagiat* erfassten Daten sind die die `rsIdr`[3]-Metadaten[4]. Das Textverarbeitungsprogramm (z.B. Microsoft Office) wählt bei jedem Speichervorgang einen zufälligen Wert, mit dem alle seit dem letztem Speichervorgang hinzugefügten, entfernten oder bearbeiteten[5] Textteile markiert werden. Mit der `rsIdr` lassen sich Aussagen über die Länge der einzelnen Arbeitssitzungen und über das Einfügen von Texten aus anderen Dokumenten machen. Allerdings ist keine zeitliche Zuordnung oder Ordnung der Arbeitssitzungen möglich.

Außerdem erlauben Anzahl der Speichervorgänge, Erstellungs- und das Bearbeitungsdatum eine Einordnung der `rsIdr`-Informationen. Die nachfolgende Tabelle zeigt, welche Auswirkungen häufige Operationen auf diese Daten (mit Microsoft Office) haben:

Aktion	#Speichervorgänge	`rsIdr`	Erstellungsdatum	Bearbeitungsdatum
Erstellen einer Datei	=1	-	Wird gesetzt	Wird gesetzt
Hinzufügen/ Ändern von Text	-	rsIdr-Zuweisungen	-	-
Speichern	+1	neue rsIdrs	-	Wird gesetzt

1. formal Office Open XML Document
2. formal Microsoft Word Binary
3. Section Addition Revision ID / Revision Identifier for Paragraph
4. Dokumentiert in ECMA-376 4th edition Part 1, § 17.15.1.70
5. (aber nicht anders formatierten)

Aktion	#Speichervorgänge	`rsIdr`	Erstellungsdatum	Bearbeitungsdatum
Speichern unter neuem Namen	=2	-	Wird gesetzt	Wird gesetzt
Einfügen von kopiertem Inhalt	-	Neuer Inhalt taucht unter **einer** rsIdr auf	-	-
Umbenennen im Explorer	-	-	-	-

Sicherheitsabschätzungen

Die Verwendbarkeit der Ergebnisse von dplagiat hängt von einer zentralen Voraussetzung ab: 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.

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 `rsIdrs` 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 sehr große Sektionen dominiert wird und nur selten gespeichert wurde, dann muss der Inhalt aus einer anderen Datei kopiert worden sein. Natürlich ist es vorstellbar, dass Studenten in der Abgabehektik das gesam-

te 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[6]; und deutet sonst auf das Verfassen des Dokuments auf einem anderem Rechner, also wahrscheinlich von einer anderen Person hin.

Fazit

Mittels dplagiat kann festgestellt werden, ob große Teile eines docx-Dokumentes aus einem anderem Dokument eingefügt wurden. Da es aber natürlich nicht feststellbar ist, *wer* das Originaldokument verfasst hat, kann eine Analyse nur Indizien, aber keine Beweise für ein Plagiat liefern.

6. z.B. durch das (atypische) Verfassen einer Arbeit in vielen Textdokumenten mit andauerndem Hin- und Her-Kopieren