Fachhochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Schlagwortgenerierung für große Dokumentenportfolios und Integration durch ein Business-Intelligence-Tool

Benjamin Arndt

Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 05.03.2015

Aufgabenstellung

Diese Arbeit befasst sich mit einer Aufgabenstellung in dem Bereich „Big Data". Für die Firma mapegy (www.mapegy.com) wurden Schlagwort-Extraktionsalgorithmen evaluiert um mit diesen die folgenden Anwendungsfälle zu bearbeiten

- Beschreibung von einzelnen Dokumenten
- Beschreibung von Dokumentenportfolios
  - Word Cloud
  - Technologielandkarte – Abb. 1

Die Umsetzung dieser Algorithmen erfolgte als RapidMiner-Operator um eine Integrierung in das System und um einen Vergleich mit den bisherigen Lösungen, falls vorhanden, zu ermöglichen.

RapidMiner

Das an der Universität Dortmund entwickelte RapidMiner ist eine Java-basierte graphische Entwicklungsumgebung, die das Arbeiten im Bereich des Data Minings ermöglicht [RK01]. Durch die „Text Mining Extension" wird es ermöglicht, dass RapidMiner effizient Text Mining Aufgaben bearbeiten kann [HK13].

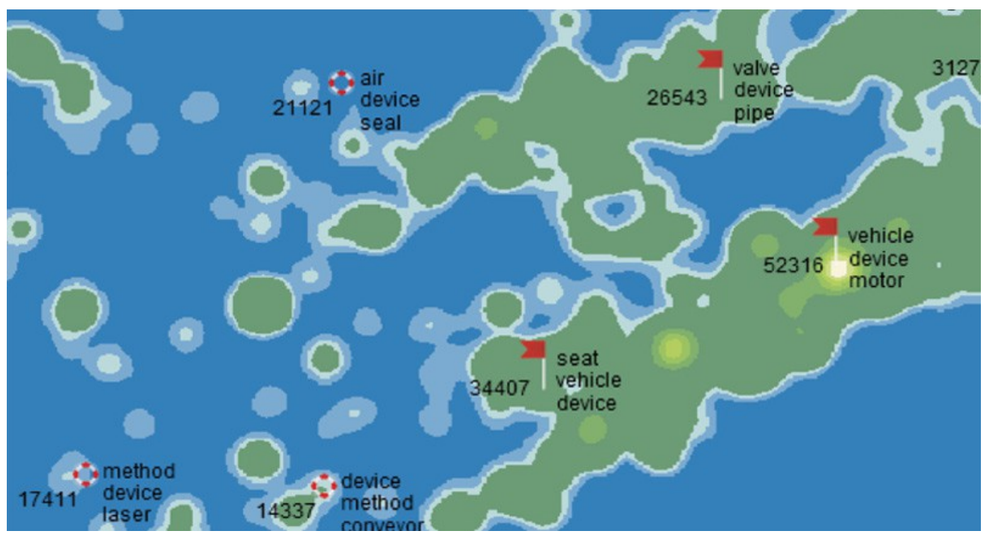

Abb. 1: Technologielandkarte (Quelle: mapegy)

Evaluierung

Die evaluierten Algorithmen lassen sich in zwei Gruppen einteilen, abhängige und unabhängige Algorithmen.

In Tabelle 1.1 (5000 Dokumente) ist die Evaluierung der abhängigen Algorithmen dargestellt, diese berechnen die Schlagwörter für einzelne Dokumente aus dem Zusammenhang im gesamten Portfolio.

Die unabhängigen Algorithmen werden in Tabelle 1.2 (1 Dokument mit 465 Wörtern) dargestellt und bei diesen werden die Schlagwörter für die Dokumente nur auf Grundlage des jeweiligen Dokumentes berechnet.

Als Ergebnis dieser Evaluierungen wurden die Algorithmen TFIDF, für die abhängigen, und TextRank, für die unabhängigen Algorithmen gewählt und als RapidMiner-Operator implementiert.

1.1	Phrasen	Zeit-Ø	Zeit-Max	Zeit-Min
TFIDF	18802	80s	138s	38s
CorePhrase	15451	6961s	12823s	5698s

1.2	Phrasen	Zeit-Ø	Zeit-Max	Zeit-Min
TextRank	77	23ms	227ms	8ms
Rake	265	120ms	866ms	52ms
Word cooc.	63	906ms	2159ms	785ms

Tab. 1: Evaluierung der Algorithmen auf einer Testmaschine mit einem AMD E2-1800 Prozessor (1,7 Gigahertz) und 16 Gigabyte Arbeitsspeicher

Implementierung

Es wurden alle evaluierten Algorithmen in Java implementiert, sowie die ausgewählten Algorithmen als RapidMiner-Operatoren. Der TFIDF Algorithmus wurde in einer erweiterten Version implementiert, diese ermöglicht die Extraktion von Schlagwortphrasen und nicht nur von einzelnen Wörtern.

Ergebnisse

Mit den erstellten Operatoren ist es möglich die Aufgabenstellungen zu bearbeiten.

Der TFIDF RapidMiner-Operator erlaubt es Schlagwörter für die Word Cloud und die Technologielandkarte zu berechnen. Ein Vergleich mit dem TFIDF-Operator, der TextMining Extension, ergab einen deutlichen Geschwindigkeitsvorteil. Dieser wurde erreicht durch das an die Problemstellung angepasste Ausgabedatenformat, für welches keine Nachbearbeitung nötig ist, erreicht.

Für den Anwendungsfall, dass einzelne Dokumente beschrieben werden, kann der TextRank RapidMiner-Operator genutzt werden. Das angepasste Ausgabedatenformat benötigt keine Nachbearbeitung und führt dazu, dass die Berechnung ohne spürbaren Zeitverlust im System durchgeführt werden kann.

Fazit

Eines der Fazite dieser Arbeit ist, dass es manchmal die einfachen Algorithmen sind, die sich durchsetzen, da sie leichter erweiterbar und anpassbar auf die spezifischen Anforderungen der Anwendungsfälle sind.

Quellen

[RK01] Ritthoff, O.; Klinkenberg, R.; Fischer, S.; Mierswa, I.; Felske, S.: *Yale: Yet another learning environment.* In LLWA 01-Tagungsband der GI-Workshop-Woche, Dortmund, Germany, pages 8492, 2001.

[HK13] Hofmann, M.; Klinkenberg, R.: *RapidMiner: Data mining use cases and business analytics applications.* CRC Press, 2013.

Betreuung Dipl.-Inform. Ingo Boersch • Fachhochschule Brandenburg • Prof. Dr. rer. nat. Gabriele Schmidt • Fachhochschule Brandenburg • Uwe Kuehn • mapegy GmbH