Themenvorschlag für eine Masterarbeit in der Web Science Group
am Lehrstuhl Informationssysteme (Prof Naumann)

NER Tinder –

Active Learning for Adaptive Ingestion Pipelines

Dieser Themenvorschlag entstand in einem attraktiven Forschungsumfeld am HPI (http://hpi.de/naumann/) und im Rahmen eines Drittmittelprojekts zur Exploration großer Datensammlungen.

Problem

- Text Mining auf Textdokumenten
- Zeichenfehler durch unsaubere Scans oder Tippfehler vom Nutzer
- Algorithmen extrahieren viele fehlerhafte Entitäten, Phrasen, oder Themen
- Automatisierte Beseitigung ohne Informationsverlust ist sehr schwierig, Manuelle Filter sind wenig flexibel

Lösungsansatz

- Entwicklung eines **NER-Tinder**s
- Named Entity Extraction ist gekoppelt an ein aktives Machine Learning Modell
- Nutzer bewerten extrahierte Entitäten während System lernt häufige Fehler zu beseitigen indem es Regeln zur Bereinigung korrupter Zeichenketten generiert
- ggf auch als human-in-the-loop Named Entity Linking System

Über das Projekt:

Nach Enthüllungen wie den Panama-Papers stehen Journalisten vor der Aufgabe große firmeninterne Datensammlungen zu durchforsten und Zusammenhänge und Fakten herauszuarbeiten. Auch Spezialisten bei Audits stehen vor dem ähnlichen Problem, dass ein Gesamtüberblick ohne langwierige Einarbeitung unmöglich ist. In diesem Forschungsprojekt werden Ansätze entwickelt, um die Kerninformationen solcher Sammlungen automatisiert zu strukturieren und visualisieren.
Das Projekt schneidet die Forschungsfelder von **Textmining**, **Dokumentenklassifikation**, **Named Entity Extraction** und **Linking**, **Relationship Extraction**, sowie die **Analyse von (sozialen) Netzwerk Graphen** und **Data Visualisation**.

Kontakt:

Ich freue mich auch über Anregungen und komplett eigene Ideen, die mit unstrukturierten Daten aus Text und Web zu tun haben.
Buzzwords: Ingestion, Extraction, Clustering, Visualisation.

Klingt interessant? Dann schreib‘ doch einfach eine E-Mail: tim.repke@hpi.de
Bei Fragen stehen wir vorab gern zur Verfügung.

Das Hasso-Plattner-Institut (www.hpi.de) ist eine in Deutschland einzigartige Forschungs- und Lehreinrichtung, die weltweit beachtete Forschung betreibt. Als Digital Engineering Fakultät der Universität Potsdam bietet das HPI innovative Studiengänge im Bereich des Digital Engineering an und betreut Promotionsprojekte. Wir verstehen uns als eine „Education Company“ und sind ob unserer Besonderheit viel beachtetes Pilotprojekt, das im Fokus von Politik und Öffentlichkeit steht.

Design IT. Create Knowledge.

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam