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

Quagga^{reloaded} –

Bringing Back Structure to Free Text Email Conversations

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

Problem

- E-Mails enthalten oft Kopien von vorangegangenen Diskussionen, semi-strukturierte Metadaten, und Boilerplates
- Text Mining auf solchen Rohtexten liefert fehlerbehaftete Ergebnisse
- Netzwerkanalyse auf E-Mail-Metadaten ist möglicherweise unvollständig, da zusätzliche Informationen im Freitext enthalten sind

Lösungsansatz

- In einem Paper haben wir „Quagga“ entwickelt, einen Deep Learning Ansatz zur Erkennung von Metadaten im Text
- Erweiterung des Systems, sodass es „lernt“ semi-strukturierte Informationen aus entsprechenden Blöcken zu extrahieren
- Erweiterung des Systems zur Erkennung von Boilerplate Texten (Signaturen, Grußworte,...)

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