Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

Frühlingssemester 2017

Bachelor-Diplomarbeit Informatik

Aufgabe für
- Andreas Waldis
- Patrick Siegfried

1. Titel (kurz und aussagekräftig, max. ca. 50 Zeichen)

Automatisierte Generierung von plattformübergreifenden Wissensnetzwerken mit Metadaten und Volltextindexierung

2. Fachliche Schwerpunkte (max. 5 Stichworte)

Text Analytics, Knowledge Engineering, Machine Learning, Artificial Intelligence

3. Einleitung (Hintergrund, Ausgangslage, Hinführung zur Problemstellung/Aufgabenstellung)

Das Forschungsprojekt Intuitive Knowledge Connectivity (IKC) entwickelt einen Prototyp zum Thema plattformübergreifendes Wissensnetzwerk. Die grundlegende Datenbank basiert auf Graphen, welche in der persönlichen Cloud gespeichert werden. Der Prototyp hat die Integration möglichst vieler Cloud-Plattformen als Ziel. Aus diesen kann der Benutzer beliebige Informationen in sein Wissensnetzwerk aufnehmen und (bisher manuell) verknüpfen. Momentan kann der Benutzer seine Informationen über eine Suchfunktion oder, falls vorhanden, hierarchische Struktur auswählen.

4. Aufgabenstellung (konkrete Aufgaben/Zielsetzungen [inhaltlich], allenfalls Hinweise zur Vorgehensweise)

Aufgrund des Inhalts und der Metadaten von Dokumenten kann der Benutzer an diversen Stellen in seinem Agieren unterstützt werden. So kann die Generierung von Wissensnetzwerken teilweise automatisiert werden. Die konkrete Problemstellung ist es, Möglichkeiten für die automatische Verknüpfung von digitalen Objekten aufzuzeigen und zu implementieren.

Eine erste, einfache Möglichkeit ist die Verknüpfung über gemeinsame Metadaten zu implementieren. Dies ist insbesondere möglich über die Zeitdimension. Dokumente können somit über ihre Zeitstempel Tagen, Wochen, Monaten und Jahren zugeordnet werden. Die Frage ist hierbei, wie die Zeitdimension abgebildet wird, d.h. ob diese ebenfalls als Netzwerk von (Tages-, Wochen-, Monats- und Jahres) Knoten abgebildet wird, ob eine Kalenderansicht eingeführt werden soll, oder ob Knoten im selben Zeitraum über beschriftete Kanten verknüpft werden (am gleichen Tag / im gleichen Monat). Weitere Möglichkeiten sind die Verwendung von Pfadinformationen in Dropbox (Ordnerstruktur) oder Tags in Evernote, welche ebenfalls Metadaten zu Dokumenten darstellen.

Eine zweite Möglichkeit ist die Verknüpfung über Analysen des Dokumentinhalts mit automatischer Schlüsselwertextraktion (Keyword Extraction). Der IKC Prototyp weist jetzt bereits die Möglichkeit zur Volltextindexierung auf, wendet diese jedoch nur für Knoten im Netzwerk an an, jedoch nicht für Dokumentinhalte auf den Cloud-Services. Werden Dokumente im Volltext indexiert, können Volltext-Engines wie Lucene oder Lunr eine sortierte Liste mit relevanten Schlüsselwörtern pro Dokument ausgeben. Dies basiert häufig auf der TF/IDF Metrik, welche die Relevanz von Suchwörtern für Dokumente im Vergleich zum Gesamtcorpus sehr gut bewertet. Auf dieser Grundlage ist es möglich,

1.) Den Volltext von Dropbox-Dateien und EvernoteNotizen in IKC zu durchsuchen

2.) Relevante Schlüsselwörter pro Dokument zu extrahieren

3.) Schlüsselwörter automatisch dem Wissensnetzwerk hinzuzufügen

4.) Dokumente mit den Schlüsselwörtern automatisch zu verknüpfen.

In dieser Diplomarbeit wird der Fokus auf die zweite Möglichkeit gesetzt. Basis dazu ist die Volltext-Indexierung von Dropbox- und Evernote-Dokumenten. Hierfür braucht es eine automatische ereignisbasierte Synchronisation mit den Cloud-Services. Darauf aufbauend können diese Dokumente (bzw. deren Metadaten) automatisch mit relevanten Schlüsselwörtern verknüpft werden. Ziel ist, dass alle Dropbox- und Evernote-Dokumente automatisch indexiert, ins Netzwerk integriert und mit Schlüsselwörtern verknüpft werden. Die Schlüsselwörter sollen ebenfalls als Knoten im Netzwerk abgebildet werden. Änderungen an den Quelldateien sollen im Wissensnetzwerk und im Index laufend synchronisiert werden.

- Automatische Abbildung von Dropbox- und Evernote Dokumenten, Ordnern, Tags als Knoten und deren Verknüpfung als Kanten im Netzwerk. Der Benutzer kann steuern / konfigurieren welche Unterordner Indexiert werden.
- Synchronisation von Änderungen (optional: automatisch)
- Automatische Volltext-Indexierung des Dokumentinhalts aller Dropbox- und Evernote-Dokumenten
- Automatische Extraktion von Schlüsselwörtern zu den Dokumenten, deren Abbildung im Netzwerk und deren Verknüpfung mit Knoten und mit Dropbox- und Evernote-Dokumenten (Automatisches Tagging im IKC)
- Optional: Abbildung einer Zeitdimension im Wissensnetzwerk und Verknüpfung von Dokumenten über Zeitstempelkategorien (wie Jahren, Monaten, Wochen und Tagen)

→ Software-Lieferobjekte:

1. Volltextindexierung von Dokument-Verzeichnissen wie Dropbox, Evernote, lokaler Client usw. (PoC: Speicherung Index & ev. Alle Metadaten in StorJ)
2. Synchronisation des Index bei Änderungen (DropBox / Evernote Events)
3. Automatische Generierung von Keywords aufgrund der TF/IDF-Methode im Gesamtcorpus (Keywords als Nodes im Wissensnetzwerk mit fixem Titel)
4. Darstellung von Keywords als Tag-Liste in den Nodes

Optional / Bonus:

1. Automatische Generierung der Zeitdimension und Verlinkung der Dokumente und Nodes anhand ihrer Daten (Zeitelemente als Nodes im Wissensnetzwerk mit fixem Titel)
2. Darstellung der Zeitdimension als Kalender-Ansicht

5. Durchführung der Arbeit

Termine

Start der Arbeit:	Montag 20. Februar 2017 (KW 08)
Zwischenpräsentation:	26.04.2016 (in der Zeit vom Montag 24. April bis Freitag 12. Mai, KW 17 bis 19)
Abgabe Schlussberichte*:	Freitag 9. Juni 2017 bis 16:00 Uhr z.H. Schulsekretariat
Abgabe Poster/WebAbstract	Termin folgt noch, Abgabe z.H. von Josef Marti
Schlusspräsentation:	21.06.2017 (in der Zeit vom Montag 19. Juni bis Donnerstag 6. Juli)
Öffentliche Diplomausstellung:	Freitag 7. Juli 2017

* Drei vollständige Schlussberichte müssen direkt z.H. des Schulsekretariats abgegeben werden. Die termingerechte Abgabe wird mit einem Stempel bestätigt.

Ein Schlussbericht inkl. elektronischem Anhang muss per Download-Link zur Verfügung gestellt werden.

Organisatorisches (Informationen zu Arbeitsplatz, Kontaktzeiten, usw.)

→

- In der ersten Woche wird die Aufgabenstellung fixiert.
- In der zweiten Woche findet das Sprint Planning gemeinsam als Workshop statt.
- Anschliessend finden 6 Sprints à 2 Wochen statt, bei denen die Planung laufend agil angepasst und Priorisiert wird.
- Es gibt wöchentliche Teammeetings. Dazu wird ein Serientermin eingerichtet.

6. Dokumentation

Für die Dokumentation gelten die grundlegenden Vorgaben des Bachelor-Studiengangs Informatik (siehe "PAWI-BDA Leitfaden Abteilung I").

Der Schlussbericht ist in 3-facher Ausführung zu erstellen (gegebenenfalls ein zusätzliches viertes Exemplar für den Wirtschaftspartner/Arbeitgeber; bitte rechtzeitig abklären!). Der Schlussbericht enthält insbesondere

- Titelblatt gemäss offizieller Vorgabe (folgen noch).
- Rückseite des Titelblattes mit folgender Selbstständigkeitserklärung:
  "Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche verwendeten Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.
  Rotkreuz, Datum, eigenhändige Unterschrift"
- Abstract in Englisch mit maximal 2'000 Zeichen.

Zusätzlich muss dem betreuenden Dozenten / der betreuenden Dozentin ein Download-Link mit dem Bericht (inkl. Anhänge), mit den Präsentationen, Messdaten, Programmen, Auswertungen, usw. abgeben werden.

Für die öffentliche Diplomausstellung ist ein Poster und für die Website ist ein Web-Abstract gemäss Vorgaben von Josef Marti zu erstellen.

→ Erarbeiten Sie für den Schlussbericht eine einheitliche Struktur in einem einzigen Dokument. Verzichten Sie so weit wie möglich auf Anhänge. Hier ist ein Vorschlag für die Inhaltsstruktur:

1. *Einleitung*: Problemstellung und Ziel der Arbeit;

2. *Stand der Technik*: Verwandte Arbeiten und Technologien, welche verwendet werden oder hilfreich sind
3. *Lösungsdesign*: Projektplanung; Darstellung der Anforderungen; Darstellung der Datenstrukturen; Beschreibung und Diskussion der möglichen Lösungsvarianten (Konzepte)
4. *Implementation*: Beschreibung des Prototyps; Benutzerhandbuch
5. *Evaluation*: Vergleich der implementierten Lösung mit der Problemstellung und den Anforderungen; Diskussion der Vor- und Nachteile; ev. Befragung zur Kundenzufriedenheit
6. *Schlussfolgerungen*: Zusammenfassung; Erkenntnisse; Potenzial; Ausblick
7. *Literatur*: Quellen und verwendete Arbeiten, Texte, Software, Links, …
8. *Anhang*: Z.B. ProjektmanagementSitzungsprotokolle, …

7. Fachliteratur/Web-Links/Hilfsmittel (optional)

8. Zusätzliche Bemerkungen (optional; spezielle Vereinbarungen, Geheimhaltungserklärung, Intellectual Property, usw.)

9. Industrie-/Wirtschaftspartner bzw. Arbeitgeber (Kontaktpersonen)

10. Verantwortlicher Dozent / verantwortliche Dozentin, Betreuungsteam

Prof. Dr. Michael Kaufmann

11. Experte/Expertin

Urs Zumstein
urs.zumstein@amanox.ch
079 639 42 58
Amanox Solutions AG
Falkenplatz 11
3012 Bern

12. Beilagen

- Bewertungsraster für Bachelor-Diplomarbeit

(verantwortlicher Dozent / verantwortliche Dozentin)

Ort, Datum, Unterschrift