Text Mining mit Open Semantic (Desktop) Search – eine digitale Such- und Annotationsumgebung für informationsgetriebene Fragestellungen in den Geisteswissenschaften.

Wettlaufer, Jörg
jwettla@gwdg.de
Georg-August-Universität Göttingen, Deutschland

Ziehe, Stefan
stefan.ziehe@stud.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Mandalka, Markus
info@opensemanticsearch.org
Freier Journalist und Informatiker, Berlin

Einleitung

In den Geisteswissenschaften sind wir bei vielen Fragestellungen auf die Sammlung und das Zusammenstellen von (teilweise sehr großen) Informationsmengen angewiesen, aus denen relevante Details herausgefiltert und anschließend in neue Zusammenhänge gesetzt werden müssen (Wettlaufer 2016). Dieser informationsgetriebene Ansatz geisteswissenschaftlichen Arbeitens ist nicht weit von den Aufgaben entfernt, die Journalisten heute für ihre Arbeit im digitalen Zeitalter lösen müssen (z.B. die Auswertung der sog. „Panama Papers“). In diesem Kontext wurde „Open Semantic Desktop Search“ (OSDS) von Markus Mandalka (Berlin) entwickelt, ein Recherche Tool bzw. eine Enterprise Suchmaschine für die semantische und explorative Suche in großen Dokumentenmengen, das auch Textanalyse und Text Mining mit offenen Standards für Linked Open Data bietet: http://www.opensemanticsearch.org . In diesem Workshop werden einige Funktionen und Anwendungsbeispiele des Softwarepakets in den digitalen Geisteswissenschaften vorgestellt sowie Hinweise zur Installation und Verwendung des Tools gegeben. Ziel des Workshops ist es die TeilnehmerInnen in die Lage zu versetzen, OSDS eigenständig für die Beantwortung wissenschaftlicher Fragestellungen, z.B. aus dem Bereich der digitalen Geschichtswissenschaften, zu verwenden und für ihre jeweiligen Bedürfnisse zu konfigurieren bzw. zu erweitern.

Herausforderungen

Korpora in der informationsgetriebenen geisteswissenschaftlichen Forschung sind oft heterogen und schlecht für Werkzeuge des Natural Language Processing (NLP) aufbereitet. Sie stammen z.B. direkt aus google books oder anderen digitalen Bibliotheken, die manchmal entweder nur eine Bilddatei oder aber einen Textlayer mit einem OCR erkannten, nicht weiter strukturierten Text enthalten. Im Fall von historischen Texten oder Textsammlungen kommen auch unterschiedliche Sprachstufen, Grammatiken etc. hinzu. Für Forschende, die eine geringe Erfahrung mit NLP und den verfügbaren Werkzeugen haben (z.B. Weblicht[1], Standford NLP[2] etc.), übersteigt der Aufwand für die Aufbereitung der Texte für eine korpuslinguistische Analyse oft die zur Verfügung stehende Zeit, zumal die Fragestellungen eher semantischer als sprachhistorischer Art sind. Wie kann man trotzdem diese Texte linguistisch „bewusst“ nutzen und erschließen?

Ein weiterer Aspekt ist das Bedürfnis vieler WissenschaftlerInnen, eine digitale Bibliothek mit fortgeschrittenen Retrieval-Funktionen für das eigene projektspezifische Textkorpus zur Verfügung zu haben. Oft verfügen sie aber selber nicht über die notwendigen Kenntnisse, um eine solche digitale Bibliothek mit einer leistungsfähigen Suchumgebung mit Keyword-in-Context (KWIC) Funktion zu erstellen. Die üblicherweise dafür zur Verfügung stehenden Technologien (Apache SOLR oder ElasticSearch) bedürfen einer gewissen Einarbeitung, einer projektspezifischen Konfiguration und einer Serverumgebung, um die gewünschten Funktionen zur Verfügung zu stellen. Einen Weg aus dieser Situation bieten z.B. Referenzmanager, die heute teilweise Volltextsuche in Dokumenten (zumeist PDF) als Zusatzleistung mit anbieten (Zotero, Mendeley, Citavi). Hier können größere Datenmengen jedoch oft nur über kostenpflichtige Speichererweiterungen realisiert werden und eine Anbindung an eine Referenz bzw. Metadaten ist für eine zu indizierende Datei obligatorisch. Dies kann bei einer großen Sammlung von Texten bzw. einem umfassenden Korpus, für das keine (geeigneten) Metadaten zur Verfügung stehen, ein großes Problem sein. Eine weitere Erschließung der Texte mittels einer Named Entity Recognition (NER), eine facettierte Suche oder die Visualisierung der gefundenen Zusammenhänge sind dabei bislang ebenfalls nicht möglich.[3]

Open Semantic Desktop Search als integrierte Text-Mining und Retrieval-Umgebung

OSDS ist ein freies Softwarebundle, dass aus Open Source Bestandteilen zusammengestellt wurde und auf dieser Grundlage als Donationware weiterentwickelt wird. Da es sich um ein Softwarepaket handelt, das üblicherweise in Serverumgebungen läuft, wird OSDS lauffertig als virtuelle Maschine (VM) angeboten, die mit Virtual Box von Oracle[4] betrieben werden kann. Die VMs können dabei als Appliance in den folgenden drei Varianten verwendet bzw. eingelesen werden: 1. Open Semantic Desktop Search (OSDS); 2. Open Semantic Search (OSS); 3. Open Semantic Search Server (OSSS).[5]

Paket 1 ist eine VM Appliance, die man mit Virtual Box laden und lokal auf einem Rechner betreiben kann. Die Appliance wird in einer aktuellen Version (Juli/August 2018) in zwei Sprachvarianten angeboten: einmal mit englischen und einmal mit deutschen Keyboard Settings. Die zweite Variante ist ebenfalls eine Appliance, die unter Oracle Virtual Box läuft, aber nur einen Server als „localhost“ bereitstellt. Dort fehlt der Gnome Desktop im Debian Linux, auf das alle Distributionen aufsetzen. Die OSDS Version schlägt mit 3GB zu Buche, die Servervariante OSS mit (nur) 1.8 GB. Das dritte Paket (OSSS) ist mit etwa 300 MB am Leichtgewichtigsten, aber erwartet natürlich auch eine (manuelle) Installation und vor allem Konfiguration auf einem Debian oder Ubuntu basierten Server. Vor allem OSDS und OSS benötigen eine moderne Hardware und ausreichend Arbeitsspeicher, um die Indexierung der Texte und die NER Pipeline schnell zu erledigen. Mit 8GB Ram und mind. einem Zweikernprozessor sollten diese Voraussetzungen aber bei den meisten TeilnehmerInnen des Workshops gegeben sein. Im Fall von diesbezüglichen Problemen ist geplant, in Zweiergruppen zu arbeiten.

Kernstück der Enterprise Suchmaschine ist ein Lucene SOLR Indexer, mit dem aufgrund der Einbindung von Apache Tika[6] recht beliebige Dokumente indexiert werden können. Die enthaltenen Informationen werden damit als Keyword im Kontext find- und referenzierbar. In dem Paket ist auch ein sogenanntes Open Semantic ETL (Extract-Transform-Load) Framework integriert, in dem die Texte für die Extraktion, Integration, die Analyse und die Anreicherung vorbereitet werden.[7] Es handelt sich um eine Pipeline, die einem sehr viel Arbeit hinsichtlich der Bereitstellung der Texte für den Indexer abnimmt. OSDS übernimmt automatisiert die Aufbereitung der Texte und kümmert sich nach dem Prinzip eines überwachten Ordners mit einer NLP Pipeline um sämtliche Schritte, die von der Extraktion über die linguistische Analyse bis zur Anreicherung mit weiteren Metadaten notwendig sind. Schließlich stellt OSDS auch einen Webservice (Rest-API) für die maschinelle Kommunikation sowie ein durchdachtes User Interface (Django[8]) zur Verfügung, mit dem die Suchmaschine bedient, konfiguriert und natürlich auch die Daten durchsucht werden können. Die facettierte Suche spielt dabei eine besondere Rolle, da die Facetten mehr oder weniger automatisch aus der linguistischen Analyse der Texte und auf der Grundlage von (konfigurierbaren) Namen Entities (Personen, Orte, Organisationen etc.) gebildet werden. Entsprechend sind auch die Hauptfunktionen des Softwarepakets angelegt: Suchinterface, ein Thesaurus für Named Entities (auch über SPARQL z.B. aus wikidata integrierbar), Extraktion von Entitäten in neu zugefügten Texten, Laden von Ontologien (z.B. in SKOS), eine listenbasierte Suche sowie eine Indexfunktion, die den Aufbau eines neuen Suchindex anstößt. Datenquellen können entweder lokale Ordner oder Ressourcen im Internet (Webseiten, Datenbanken etc.) sein. Diese können in konfigurierbaren Zeitabständen überprüft und neue Informationen dem Index hinzugefügt werden.

Lernziele und Ablauf

Durch den Workshop sollen die TeilnehmerInnen in die Lage versetzt werden, die OSDS Such- und Text Mining Funktionen in einer lokalen Umgebung in Betrieb zu nehmen, zu konfigurieren, Daten in den Suchindex einzulesen und Suchanfragen auszuführen. Dabei soll die Verwendung des NER-Taggers und die Nutzung der Thesaurus-Funktion für die Bereitstellung eigener Suchfacetten erklärt und die Benutzung anhand praktischer Beispiele eingeübt werden. Nach der Erprobung verschiedener Suchanfragen zum Beispielkorpus werden die erweiterten Funktionalitäten von OSDS vorgestellt und getestet. Eine wichtige zusätzliche Funktionalität betrifft die integrierte OCR mit tesseract,[9] die auch Frakturschriften erkennen kann. Desweiteren gibt es eine Schnittstelle der NER zur neo4j Graphdatenbank,[10] die eine Darstellung der Entitäten und ihrer Verknüpfungen als Graphen ermöglicht. Weitere Auswertungsmöglichkeiten eröffnen die Visualisierungsoptionen, die sowohl Ortsnamen auf einer Karte als auch Netzwerke zwischen Personen anzeigen können.

Der Ablauf ist wie folgt vorgesehen:
Block I: 90 Minuten Workshop

- 15 Minuten Begrüßung / Einführung
- 30 Minuten Hands On - Installation von Virtual Box / OSDS auf eigenen Geräten, sofern dies noch nicht vorher durchgeführt wurde. Behebung von Problemen.
- 30 Minuten Lecture zur Funktionalität von OSDS / Parallel: Indexierung des Beispielkorpus
- 15 Minuten Hands On / Ausprobieren des UI / Suchinterface anhand einfacher Abfragen mit boolschen Operatoren, Einsatz der Suchfacetten

30 Minuten Pause
Block II. 90 Minuten Workshop

- 30 Minuten Vorstellung der erweiterten Funktionen von OSDS (OCR, Semantic Tagging, NER, Suche mit Listen, neo4j, Visualisierungen)
- 30 Minuten Hands On: Bearbeitung verschiedener Aufgabenstellungen zu diesen Funktionen (Semantic Tagging, Einbindung von wikdata über SPARQL in die NER)
- 30 Minuten gemeinsame Diskussion der Erfahrungen mit OSDS, Ausblick des Entwicklers auf weitere Funktionalitäten und zukünftige Entwicklungen sowie Fragen.

Beitragende (Kontaktadressen und Forschungsinteressen)

Der Workshop wird angeboten von Mitarbeitern des Göttingen Centre for Digital Humanities an der Georg-August Universität Göttingen in Zusammenarbeit mit dem Entwickler von OSDS. Das Zentrum hat u.a. einen Forschungsschwerpunkt im Bereich des Text Mining und der digitalen Geschichtswissenschaft. Der Workshop wird ohne besondere Voraussetzungen angeboten (man beachte aber die Anforderungen an die Hardware) und steht TeilnehmerInnen ohne spezielle Programmier- oder Softwarekenntnisse offen. In den Hands-On Teilen wird allerdings individuelle Unterstützung bei der Installation und Einrichtung der notwendigen Virtualisierungsumgebung sowie des Enterprise Suchsystems gegeben.

Jörg Wettlaufer
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen

Die vielfältigen Forschungsinteressen von Jörg Wettlaufer liegen u.a. im Bereich Information Retrieval, Semantic Web Anwendungen in den Geisteswissenschaften und Digitale Geschichtswissenschaft. Er tritt für einen intensiven Dialog zwischen Fachwissenschaften und den Digital Humanities ein und ist seit 2018 zusätzlich im Bereich der Digitalen Lehre in den Geisteswissenschaften tätig.

Stefan Ziehe
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen

Stefan Ziehe ist Masterstudent am Institut für Informatik und zugleich wissenschaftliche Hilfskraft am GCDH und interessiert sich insbesondere für Deep Learning-Verfahren im Bereich der Sentiment Analysis und des NLP.

Markus Mandalka
Warschauerstr. 66
10243 Berlin

Die Expertise von Markus Mandalka liegt in der Schnittstelle zwischen Informatik und Journalismus. Er ist freier Journalist sowie Informatiker und Entwickler des Open Semantic Desktop Search Systems. In diesem Zusammenhang bietet er auch Beratungsservice für Institutionen und Projekte an.

Zahl der möglichen Teilnehmer: 5-25

Benötigte technische Ausstattung:

Es wird außer einem Beamer und WLAN keine besondere technische Ausstattung benötigt. Es sollte sich allerdings um Räumlichkeiten handeln, die eine aktive Betreuung der TeilnehmerInnen in den Hands-On Phasen erlaubt. Die TeilnehmerInnen müssen zudem geeignete Hardware für die praktischen Übungen selber mitbringen (Notebooks mit Prozessoren mit Virtualisierungsbefehlssätzen und 8 GB RAM).

Fußnoten

1. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.
2. https://stanfordnlp.github.io/CoreNLP/.
3. Vgl. zur semantischen Annotation aber INCEPTION: https://inception-project.github.io// und https://recogito.pelagios.org/
4. https://www.virtualbox.org/.
5. https://www.opensemanticsearch.org/de/download.
6. https://tika.apache.org/.
7. https://www.opensemanticsearch.org/etl.
8. https://www.djangoproject.com/.
9. https://github.com/tesseract-ocr/.
10. https://neo4j.com/.

Bibliographie

Gahlke, Lukas / Mai, Florian / Schelten, Alan / Brunsch, Dennis / Scherp, Ansgar (2017): *Using Titles vs. Full-text as Source for Automated Semantic Document Annotation*, in: Computing Research Repository. https://arxiv.org/abs/1705.05311 [zuletzt abgerufen 12.10.2018].

O'Caroll, Ultan (2016): *Open Semantic Desktop Search – good but…*, https://uoccou.wordpress.com/2016/04/22/open-semantic-desktop-search-good-but/ [zuletzt abgerufen 12.10.2018].

Sporleder, Caroline (2010): *Natural Language Processing for Cultural Heritage Domains*, in: Language and Linguistics Compass, 4:9, 750–768.

Wettlaufer, Jörg (2016): *Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern*, in: Zeitschrift für digitale Geisteswissenschaften. text/html Format. DOI: 10.17175/2016_011 [zuletzt abgerufen 12.10.2018]

Wettlaufer, Jörg (2018): *Hands-On „Open Semantic Desktop Search" (OSDS)*, Einstündiger Workshop auf dem Historikertag Münster am 27.09.18. Vgl. http://digigw.hypotheses.org/2475 und http://digihum.de/blog/2018/09/18/hands-on-open-semantic-desktop-search/ [zuletzt abgerufen 12.10.2018].