Projektskizze Semesterprojekt CAS Practical Machine Learning

Data Mining & NLP in e-rara

Kathi Woitas

1. Ausgangslage

E-rara[1] ist eine nationale Web-Plattform, die von wissenschaftlichen Bibliotheken gemeinsam betrieben wird, um digitalisierte Druckwerke allgemein zugänglich zu machen. Bei den Digitalisaten handelt sich um Druckwerke aus der Schweiz vom 15. bis in das 20. Jahrhundert, zudem um thematische Sondersammlungen. Die Universitätsbibliothek Bern (UB Bern) stellt hier ihre Sammlungen „Rossica Europeana", „Russisches Exilschrifttum" sowie Bernensia unter freier Lizenz zur Verfügung. Digitalisate werden dabei nicht nur als Image/PDF, sondern wo möglich bzw. vorhanden, auch als Volltextdatei (TXT) zur Verfügung gestellt. E-rara verfügt über eine OAI-PMH[2]-Schnittstelle, die ausführliche Metadaten in XML-basierten bibliothekarischen Formaten ausliefert, aber keine öffentliche Dokumentation aufweist. Ein Batch-Download ist weder für Meta- noch für Volltextdaten vorgesehen.

2. Zielsetzung

Mit der Projektarbeit soll die Nachnutzbarkeit der Digitalisierungsdaten auf e-rara untersucht und verbessert werden. Einfache Hilfsmittel zur Nutzung von Metadaten und Volltextdaten sollen entwickelt und bereitgestellt werden. Dies kann in folgende Fragestellungen gegliedert werden:

1. Datenzugang: Wie können Meta- und Volltextdaten im Batch nach bestimmten Kriterien einfach heruntergeladen werden?
2. Data Preprocessing: Welche Datentransformationen sind notwendig, um aus den Rohdaten Korpora zu erstellen, die für wissenschaftliche Analysen verwendet werden können? Welche Hindernisse bestehen bezüglich der Datenqualität?
3. Anwendung von NLP-Methoden: Welche Methoden des Natural Language Processing können auf den Korpora angewendet werden? Welche Erkenntnisse lassen sich u.U. hierbei gewinnen?
4. Implementierung: Wie können die entwickelten Methoden als niedrigschwelliges Angebot für die Benutzenden der UB Bern bzw. die Allgemeinheit zur Verfügung gestellt werden?

3. Methoden

Für die Fragestellungen werden die im CAS PML erlernten Methoden zur Programmierung mit Python und zum NLP angewendet:

1. Datenzugang: Programmierung von einfachen Datenzugängen zu spezifischen Sammlungen.
2. Data Preprocessing: Erstellung von Skripten zur Normalisierung, Tokenisierung, Lemmatisierung, Part-of-Speech-Tagging
3. Anwendung von NLP-Methoden: Exploration von verschiedenen NLP-Methoden auf den Korpus.
4. Implementierung: Bereitstellung der entwickelten Skripte mit Dokumentation zur Nachnutzung (Github).

4. Personen

Studierende	Universitätsbibliothek Bern; Woitas, Kathi; Tel. +41788994301, kathi.woitas@students.bfh.ch
Ansprechpartner / Betreuer in der Firma:	keiner

[1] e-rara: https://www.e-rara.ch/
[2] OAI Protocol for Metadata Harvesting: http://www.openarchives.org/OAI/openarchivesprotocol.html