Thema1_Datenanalyse_NLP_VersionA_Studiere 22.01.26

Masterarbeit: Entwicklung einer NLP-basierten Pipeline zur automatisierten Analyse und Extraktion von Kompetenzprofilen aus Kursbeschreibungen

Betreuer: Adrian Vogler, M.Sc., Prof. Dr. Matthias Hemmje

Projektbezug: HR-QDE (Human Resource Qualification Development Ecosystem)

Thema: 1 von 3 (Datenanalyse & NLP)

Notwendige Kenntnisse und Voraussetzungen

Für eine erfolgreiche Bearbeitung der Aufgabenstellung sind Vorkenntnisse in folgenden Bereichen notwendig bzw. nützlich:

Methodisch:

- Natural Language Processing (NLP) (notwendig)
- Information Extraction (notwendig)
- Named Entity Recognition (NER) (notwendig)
- Text Mining (nützlich)
- Machine Learning (nützlich)

Technisch:

- Python (notwendig)
- NLP-Bibliotheken (notwendig)
  - spaCy oder NLTK
  - BERT, transformers
  - scikit-learn
- Web Scraping (notwendig)
  - BeautifulSoup, Selenium
  - Requests, Scrapy
- Datenverarbeitung (notwendig)
  - pandas, numpy
  - JSON, CSV, XML
- Ontologien und semantische Modellierung (nützlich)
  - RDF, OWL
- Git (notwendig)

Motivation und Ausgangspunkt: Der Studierendenkompass als Navigationskonzept

Studierende in grundständigen Studiengängen stehen häufig vor einem komplexen Navigationsproblem. Bildlich gesprochen gleichen sie Weltseglern, die zwar das Handwerk der Navigation beherrschen und über Seekarten (Curricula, Modulhandbücher) verfügen, deren konkreter Zielhafen (berufliche Qualifikation) jedoch oft im Nebel liegt. Während sie lernen, wie man segelt (Studium durchlaufen), bleiben die spezifischen Anforderungen der Industrie und

des Arbeitsmarktes (das Ziel) oft intransparent. Es fehlt an Instrumenten, die nicht nur eine Navigation durch das Studienprogramm ermöglichen, sondern diese Navigation gezielt auf aktuell nachgefragte Berufsbilder ausrichten.

Vor diesem Hintergrund wird an verschiedenen Hochschulen das Konzept eines sogenannten „**Studierendenkompasses**" diskutiert. Ein solches System soll als Orientierungshilfe dienen, um individuelle Lernpfade so zu gestalten, dass sie die Lücke zwischen dem akademischen Angebot und den realen Bedarfen des Arbeitsmarktes schließen. Ziel ist es, die Navigation vom Start (Studienbeginn) bis zum Ziel (Berufsfähigkeit in einem spezifischen Feld) durch semantische Technologien und KI-gestützte Pfadempfehlungen transparent und steuerbar zu machen.

Die Betrachtung erfolgt anhand zweier beispielhafter Institutionen, die sich in ihrer Ausrichtung stark unterscheiden:

FernUniversität in Hagen (FUH): Als größte Fernuniversität Deutschlands spielt die FernUniversität eine zentrale Rolle bei der Ausbildung qualifizierter Fachkräfte, die für die Herausforderungen der Digitalisierung und Industrie 4.0 gerüstet sind. Studierende im Fernstudium benötigen eine starke Orientierungshilfe, um individuelle Lernpfade zu gestalten, da sie oft aus diversen beruflichen Kontexten kommen oder auf unterschiedliche Qualifikationsziele hinsteuern. Für diese deutschsprachigen Studierenden aus verschiedenen Ländern ist der Studierendenkompass ein Instrument zur **individuellen Orientierung** im Studium.

Hochschule Darmstadt (HDA): Die Hochschule Darmstadt ist als Teil der **European University of Technology Alliance (EUT+)** in einen europäischen Hochschulverbund eingebunden. Im Rahmen des **Bologna-Prozesses** erkennen die Partnerhochschulen von EUT+ ihre Studienleistungen gegenseitig an, was **mehrsprachige Studienprogramme** und **grenzüberschreitende Mobilität** ermöglicht. Studierende müssen verstehen, welche Kurse im Ausland ihren heimischen Anforderungen entsprechen und wie diese anerkannt werden. Hier ist der Studierendenkompass ein Instrument zur **internationalen Mobilität** und zur Sicherstellung der Vergleichbarkeit von Studienleistungen über Ländergrenzen hinweg.

Obwohl die Motivationen unterschiedlich sind (individuelle Orientierung im Fernstudium vs. internationale Mobilität im Präsenzstudium), ist die **technische Anforderung identisch**: Es wird eine **maschinenlesbare, semantische Repräsentation** von Qualifikationen und **Kompetenzen im Sinne nachgewiesener Qualifikationen** benötigt, um diese Navigation automatisiert zu unterstützen.

Gesellschaftlicher Kontext und Relevanz

Von diesem Ausgangspunkt der beiden Hochschulen aus muss der übergeordnete gesellschaftliche Kontext betrachtet werden. Obwohl die FernUniversität in Hagen und die Hochschule Darmstadt unterschiedliche Zielsetzungen verfolgen, operieren beide im gleichen sozioökonomischen Umfeld, das zunehmend durch einen gravierenden Fachkräftemangel geprägt ist (bedingt u.a. durch den demografischen Wandel).

Daraus ergibt sich für beide Institutionen die Notwendigkeit, Studierende in den grundständigen Studiengängen wesentlich gezielter und passgenauer auszubilden. Das Ziel muss sein, "brauchbare Human Resources" bereitzustellen, die dem Arbeitsmarkt ohne lange Einarbeitungszeiten zur Verfügung stehen. Um dies zu gewährleisten, gewinnen integrierte Konzepte wie das im Folgenden beschriebene Human Resource Qualification Development Ecosystem (HR-QDE) für Hochschulen massiv an Bedeutung.

Kontext: Vom Fachkräftemangel zur Qualifikationsentwicklung

Die Herausforderungen und Motivationen für die Entwicklung eines Human Resource Qualification Development Ecosystem (HR-QDE) lassen sich in zwei zentrale Perspektiven unterteilen: **Nachfrageseite (Demand-Side)** und **Angebotsseite (Supply-Side)**. Die Nachfrageseite, primär repräsentiert durch die Industrie, benötigt Fachkräfte mit spezifischen **Qualifikationen im Sinne nachgewiesener Kompetenzen**, die auf ihre Prozesse und Ressourcen zugeschnitten sind. Auf der anderen Seite können die Anbieter, wie Universitäten und Bildungseinrichtungen, durch **maschinenlesbare semantische Modelle** qualifikationsorientierte Bildungsangebote bereitstellen.

Diese klare Unterscheidung zwischen den beiden Perspektiven ermöglicht eine präzisere Adressierung von Nachfrage und Angebot. Sie dient als Grundlage für die Entwicklung eines integrierten Modells, das die Bedürfnisse der Nachfrager mit den Bildungsressourcen effektiv in Einklang bringt, um dem Fachkräftemangel nachhaltig zu begegnen.

Relevante Vorarbeiten

In diesem Kontext ordnet sich diese Arbeit in die bestehenden Forschungsaktivitäten zur Entwicklung des HR-QDE-Konzepts ein. Für diese Masterarbeit sind insbesondere folgende Vorarbeiten relevant:

SQW (Semantic Qualification Web): Die Dissertation von Lothary zum SQW beschreibt ein System, das auf Basis semantischer Technologien und KI die Generierung personalisierter, lebenslanger Lernpfade unterstützt. Das zentrale

Element von SQW ist der **Semantic Triple Creation Algorithm**, der unstrukturierte Kursbeschreibungen in maschinenlesbare RDF-Formate überführt. Dieser Algorithmus verwendet eine **NLP-Pipeline** mit spaCy für Named Entity Recognition (NER) und **Kompetenzextraktion** aus Textbeschreibungen. Diese Vorarbeit bildet die methodische Grundlage für die vorliegende Masterarbeit.

ESCO (European Skills, Competences, Qualifications and Occupations): ESCO ist ein mehrsprachiges europäisches Klassifikationssystem für Fähigkeiten, Kompetenzen, Qualifikationen und Berufe. ESCO bietet ein standardisiertes Vokabular, das für die Annotation extrahierter Kompetenzen verwendet werden kann.

Forschungslücken und Motivation für diese Arbeit

Trotz der umfangreichen Vorarbeiten im Bereich SQW bestehen spezifische Forschungslücken im Bereich **Datenanalyse & NLP** für grundständige Hochschulbildung:

1. **Fehlende Parsing-Methoden für strukturierte Hochschulkurskataloge:** Hochschulkurskataloge (HDA, FUH) liegen in unterschiedlichen Formaten vor (PDF, HTML, strukturierte Datenbanken). Es fehlen robuste Parsing-Methoden, die diese heterogenen Datenquellen einheitlich verarbeiten können.
2. **Unzureichende NLP-Methoden für Kompetenzextraktion aus Hochschulkursbeschreibungen:** Die in SQW entwickelten NLP-Methoden fokussieren auf Weiterbildungskurse. Hochschulkursbeschreibungen haben eine andere Struktur (Learning Outcomes, Modulabhängigkeiten, ECTS) und erfordern angepasste Extraktionsmethoden.
3. **Fehlende Integration von Stellenanzeigen als Demand-Side Datenquelle:** Während SQW primär auf Supply-Side (Kursangebote) fokussiert, fehlt eine systematische Methode zur Extraktion von Kompetenzanforderungen aus Stellenanzeigen (Demand-Side).

Einordnung der Aufgabenstellung

Diese Masterarbeit leistet einen Beitrag zur Entwicklung eines übergeordneten Qualifikations-Ökosystems. Sie fokussiert sich auf die **Datenanalyse und NLP-basierte Extraktion** von Kompetenzen aus Kursbeschreibungen und Stellenanzeigen und bildet damit die Datengrundlage für weiterführende Arbeiten zur semantischen Modellierung und zum Matching.

Aufgabenbeschreibung

Das Ziel dieser Masterarbeit ist die Entwicklung einer **NLP-basierten Pipeline zur automatisierten Analyse und Extraktion von Kompetenzprofilen** aus zwei Datenquellen:

1. **Supply-Side:** Kursbeschreibungen von HDA (Fachbereich Wirtschaft) und FUH (Fachbereich Informatik)
2. **Demand-Side:** Stellenanzeigen von Unternehmen (z.B. Siemens, SAP, Bosch)

Die Arbeit soll folgende Forschungsfragen beantworten:

Forschungsfragen:

1.1. Wie können Modulhandbücher von HDA und FUH automatisiert heruntergeladen (Wget), in Markdown konvertiert (Docling) und für NLP-Analyse vorbereitet werden?

- Welche Formate liegen vor (PDF, HTML, Datenbank)?
- Welche Parsing-Methoden sind geeignet (BeautifulSoup, PyPDF2, APIs)?
- Wie können Metadaten (ECTS, Semester, Voraussetzungen) extrahiert werden?

1.2. Wie können Kompetenzen im Sinne nachgewiesener Qualifikationen automatisch aus Kursbeschreibungen extrahiert werden (NLP)?

- Welche NLP-Methoden sind geeignet (spaCy, BERT, transformers)?
- Wie können Learning Outcomes identifiziert werden?
- Wie können Kompetenzbegriffe von anderen Begriffen unterschieden werden?

1.3. Wie können Stellenanzeigen von Unternehmen gecrawlt und analysiert werden?

- Welche Datenquellen sind geeignet (LinkedIn, Indeed, Unternehmenswebseiten)?

- Wie können Stellenanzeigen automatisch gecrawlt werden (Selenium, APIs)?
- Wie können Kompetenzanforderungen aus Stellenanzeigen extrahiert werden?

Phasen der Arbeit

Die Arbeit gliedert sich in vier Phasen nach der Forschungsmethodik von Nunamaker:

Phase 1: Beobachtung (Analyse und Recherche)

Ziel: Verstehen der Datenquellen und NLP-Methoden

Aktivitäten:

- Literaturrecherche zu NLP, Information Extraction, Named Entity Recognition, Text Mining
- Analyse der Datenquellen: HDA Kurskatalog, FUH Kurskatalog, Stellenanzeigen
- Identifikation von Parsing-Anforderungen

Ergebnis:

- Literaturübersicht zu NLP-Methoden für Kompetenzextraktion
- Dokumentation der Datenquellen (Formate, Strukturen, Beispiele)
- Anforderungskatalog für Parsing und Extraktion

Phase 2: Theoriebildung (Konzeption und Modellierung)

Ziel: Konzeption der Parsing- und Extraktionsmethoden

Aktivitäten:

- Konzeption der Parsing-Architektur für HDA und FUH Kurskataloge
- Design von Web-Scrapern für Stellenanzeigen
- Konzeption der NLP-Pipeline (spaCy vs. BERT vs. transformers)
- Design der Kompetenzextraktion (NER, Pattern Matching, ML)
- Definition von Kompetenz-Kategorien (Hard Skills, Soft Skills, Tools, Methoden)
- Design des Output-Formats (JSON, CSV, strukturierte Datenbank)

Ergebnis:

- Architekturdiagramm der Parsing- und Extraktionspipeline
- Spezifikation der NLP-Methoden
- Datenmodell für extrahierte Kompetenzen

Phase 3: Systementwicklung (Implementierung)

Ziel: Implementierung der Parsing- und Extraktionssysteme

Aktivitäten:

- Implementierung der Parser (HDA-Parser, FUH-Parser)
- Implementierung des Stellenanzeigen-Crawlers (Selenium, BeautifulSoup)
- Implementierung der NLP-Pipeline (spaCy-basierte NER, Pattern Matching, BERT-basierte Klassifikation)
- Post-Processing (Bereinigung, Normalisierung, Deduplizierung)
- Aufbau der Datenbank

Ergebnis:

- Funktionsfähige Parser für HDA und FUH Kurskataloge
- Funktionsfähiger Web-Scraper für Stellenanzeigen
- NLP-Pipeline zur Kompetenzextraktion

- Datenbank mit extrahierten Kompetenzen

Phase 4: Experimentierung (Evaluation)

Ziel: Evaluation der Parsing- und Extraktionsqualität

Aktivitäten:

- Evaluation der Parser (Precision, Recall, F1-Score für Metadatenextraktion)
- Evaluation der Kompetenzextraktion (Vergleich mit manuell annotierten Ground Truth)
- Qualitative Analyse (Welche Kompetenztypen werden gut/schlecht erkannt?)
- Identifikation von Verbesserungspotenzialen

Ergebnis:

- Evaluationsbericht mit Precision, Recall, F1-Score
- Dokumentation der Fehlerquellen
- Empfehlungen für Verbesserungen

Erwartete Ergebnisse

Die Arbeit soll folgende konkrete Ergebnisse liefern:

1. **Parser für HDA und FUH Kurskataloge:** Python-basierte Tools zur automatischen Extraktion von Kursbeschreibungen und Metadaten
2. **Web-Scraper für Stellenanzeigen:** Python-basiertes Tool zum Crawling von Stellenanzeigen
3. **NLP-Pipeline zur Kompetenzextraktion:** Python-basiertes System zur automatischen Extraktion von Kompetenzen aus Kursbeschreibungen und Stellenanzeigen
4. **Extrahierte Daten und annotierter Korpus:**
   - ~50-100 Kurse von HDA (Fachbereich Wirtschaft)
   - ~50-100 Kurse von FUH (Fachbereich Informatik)
   - ~50-100 Stellenanzeigen von Unternehmen
   - Extrahierte Kompetenzen für alle Kurse und Stellenanzeigen
5. **Evaluationsbericht:** Dokumentation der Parsing- und Extraktionsqualität
6. **Wissenschaftliche Dokumentation:** Abschlussbericht (Masterarbeit) mit vollständiger Dokumentation

---

Zeitplan

Die Arbeit soll innerhalb von **6 Monaten** abgeschlossen werden:

- **Monat 1:** Phase 1 (Beobachtung)
- **Monat 2:** Phase 2 (Theoriebildung)
- **Monat 3-4:** Phase 3 (Systementwicklung)
- **Monat 5:** Phase 4 (Experimentierung)
- **Monat 6:** Abschlussdokumentation

Literatur

Wichtige Referenzen:

- Lothary, S. (2024). Semantic Qualification Web (SQW). Dissertation, FernUniversität Hagen.
- Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
- Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.).
- spaCy Documentation. https://spacy.io/

- Hugging Face Transformers. https://huggingface.co/transformers/