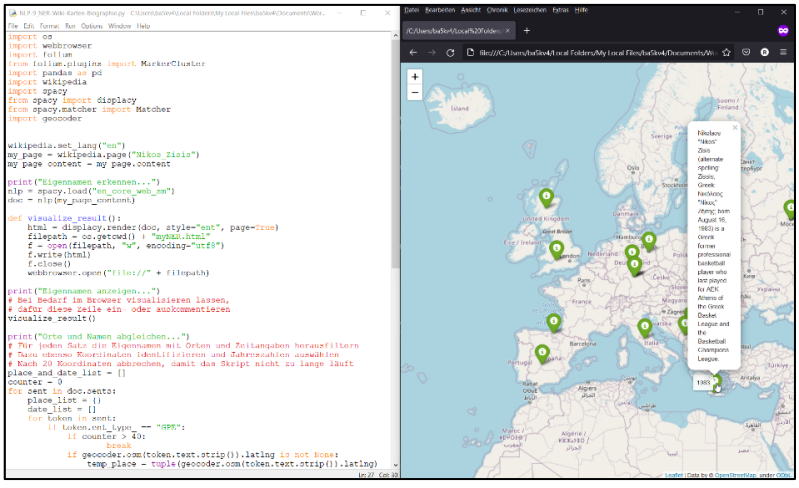
Nachbericht: Wikipedia durchforsten mit Python

Ein Angebot im Rahmen des Schülerforschungszentrums
der TechnologieAllianzOberfranken (TAO)

Im Juli 2022 fand der Workshop „Wikipedia durchforsten mit Python“ an der Otto-Friedrich-Universität Bamberg statt. Dieser Workshop war der erste in Bamberg organisierte Präsenz-Workshop seit den Einschränkungen, die durch die Corona-Pandemie ausgelöst wurden. Am Workshop nahmen 7 Schüler (Klasse 9-11) aus vier oberfränkischen Gymnasien teil, die mithilfe der Programmiersprache Python verschiedene Anwendungen im Zusammenhang mit Natürlicher Sprachverarbeitung bearbeiteten.

Der Workshop wurde konzipiert und betreut von Lutz Reuter (Kontaktlehrer des TAO-Schülerforschungszentrum) und Robin Jegan (Wissenschaftlicher Mitarbeiter am Lehrstuhl für Medieninformatik an der Universität Bamberg). Zusätzliche Unterstützung wurde durch Caroline Oehlhorn (Mitarbeiterin am TAO-Schülerforschungszentrum) angeboten, die insbesondere bei der Anmeldung und Organisation tätig war, sowie durch Felix Engl (Wissenschaftlicher Mitarbeiter am Lehrstuhl für Medieninformatik an der Universität Bamberg) in der Betreuung technischer Belange am Nachmittag.

Die 7 Teilnehmer trafen sich am 18.7. in Bamberg am ERBA-Campus der Universität auf der ERBA-Insel. In Rechnerraum des Lehrstuhls Medieninformatik wurden zunächst in einer Präsentation grundlegende Informationen zu Python vermittelt, bevor auf Bereiche der Natürlichen Sprachverarbeitung eingegangen wurde, die für den Workshop relevant waren.



Anschließend wurden den Schülern Programmierskripte übermittelt, die in zwei Schwierigkeitsgraden vorbereitet wurden, damit auch Teilnehmende mit unterschiedlich großen Vorkenntnissen in Python sowohl eine Herausforderung wie auch einen Mehrwert durch die schwierigeren Skripte erhalten konnten. In zwei Praxiseinheiten wurden die Skripte am Vormittag und am Nachmittag von den Teilnehmerinnen und Teilnehmern bearbeitet, die sie mithilfe des Python Editors IDLE (https://www.python.org/downloads/), welcher in der Standard-Installation von Python mitgeliefert wird, oder im derzeit sehr beliebten Editor Visual Studio Code (https://code.visualstudio.com/) auf den PCs des Medieninformatik-Rechnerraums ausführen konnten.

In den Skripten wurden vier Anwendungsgebiete vermittelt. Zunächst wurden Daten und Inhalte aus der deutsch- sowie englischsprachigen Wikipedia extrahiert. Daraufhin wurden zwei grundlegende Techniken der Natürlichen Sprachverarbeitung eingeführt: Zum einen die Wortarterkennung, also die Einordnung von Wörtern in Nomen, Verben, Adjektiven, etc., und zum anderen die Eigennamenerkennung, das heißt die Identifikation von Worten, die Orte, Personen, Zeitangaben oder anderen Kategorien zugeordnet werden können. Anschließend wurden diese beiden Techniken kombiniert, das heißt die aus Wikipedia gewonnenen Daten wurden mithilfe der Methoden der Natürlichen Sprachverarbeitung analysiert. Das Ergebnis dieser Kombination waren beispielsweise Frequenzlisten, das heißt eine Auflistung und Summe der verwendeten Wortarten, oder Visualisierungen, die etwa alle Orte und Personen aus einem Wikipedia-Text farbig markiert im Browser darstellen.

Als drittes Anwendungsgebiet wurden digitale Karten gewählt, die analog zu bekannten Kartenanbietern wie Google Maps oder OpenStreetMap, eine browserbasierte Ansicht auf eine Karte der Welt ermöglicht, auf der man bis auf Straßenebene hineinzoomen kann. Mithilfe der Python Skripte wurden diese Karten angelegt und Marker auf die Karte platziert. Zuletzt wurden die drei genannten Anwendungsgebiete kombiniert, um Orte und Zeitangaben aus den Wikipedia Texten herauszufiltern, die verwendet wurden, um die Stationen in der Biographie einer Person mittels Marker auf einer Karte zu platzieren. Die Schüler konnten dabei eigene Texte auswählen, diese nach unterschiedlichen Kriterien bearbeiten und die Karten und dort platzierten Marker anpassen.

Das vierte Anwendungsgebiet entstammt erneut der Natürlichen Sprachverarbeitung und umfasst die Sentiment-Analyse, das heißt die Analyse, ob ein Text eine positive oder negative Meinung enthält. Diese Aufgabe der Sentiment-Analyse wurde sowohl für Deutsch wie auch für Englisch umgesetzt und zuletzt mit dem Abruf von Inhalten aus Wikipedia kombiniert, um Wikipedia-Texte auf positive bzw. negative Meinungen zu analysieren.

Robin Jegan
(Workshop-Leiter)

Lutz Reuter
(TAO SFZ Kontaktlehrer)

Caroline Oehlhorn
(TAO SFZ Koordinatorin, Standort Bamberg)