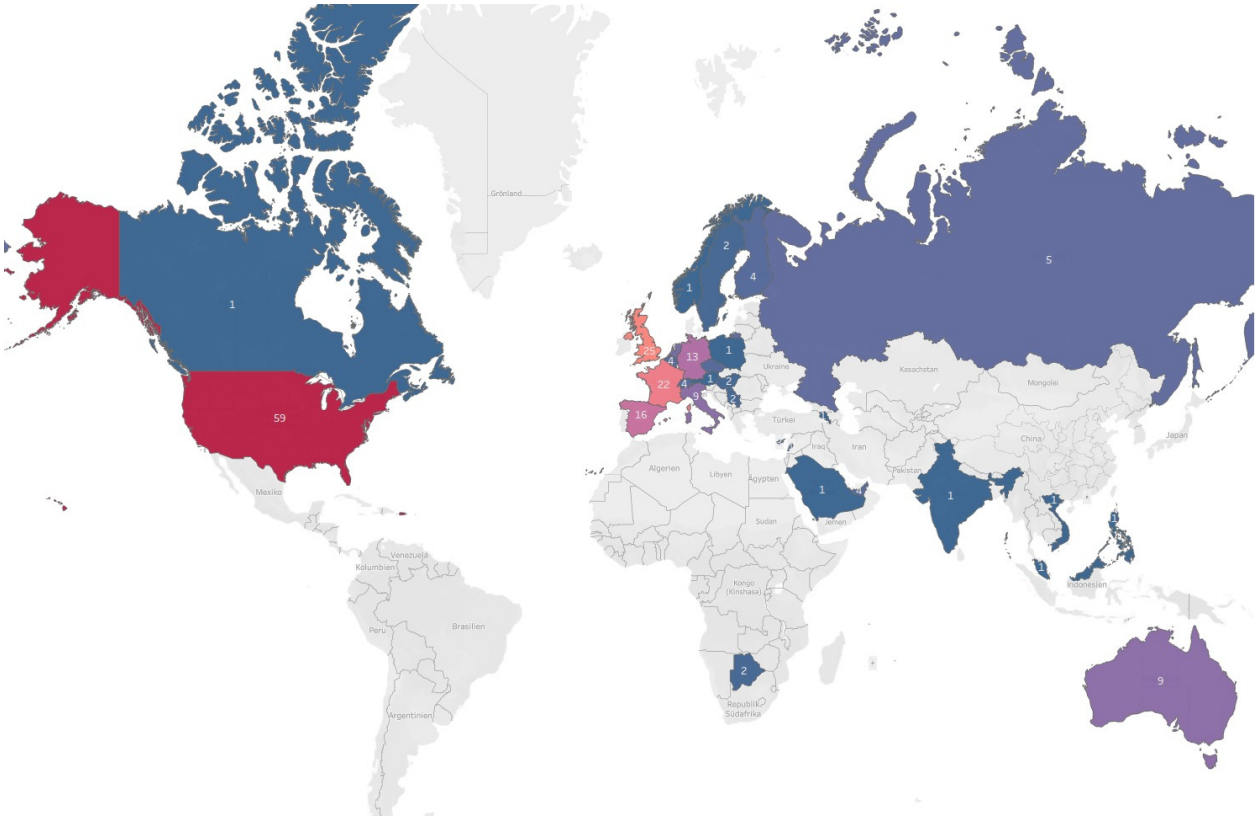
Hochschule Luzern

Master of Science in Applied Data and Information Science (MSc IDS)

Pflichtmodul «Data Collection, Integration and Preprocessing »

Erstellung einer weltweiten Studiengangs-Übersicht für die Studienrichtung “Data Science”



Verfasser

Jonas Zürcher
jonas.zuercher@stud.hslu.ch

Carole Mattmann
carole.mattmann@stud.hslu.ch

Luca Casuscelli
luca.casuscelli@stud.hslu.ch

Silvan Leibacher
silvan.leibacher@stud.hslu.ch

Betreuer

Prof. Erwin Mathis
erwin.mathis@hslu.ch

Luzern, 11.04.2020

Management Summary

Im Rahmen dieser Projektarbeit wurde zuerst theoretisches Wissen im Bereich Webcrawler, Python, Tableau Prep und Tableau erarbeitet. Aufbauend auf diesen Erkenntnissen wurde in einem mehrstufigen Prozess ein Übersichtsdatenset für alle „Data Science“ Studiengänge weltweit erstellt. Dazu wurden zuerst zwei Webseiten mit einer Auflistung von Master und Bachelor Studiengängen mit der Fachrichtung „Data Science“ und eine Webseite mit Hochschulranglisten gecrawlt. Die Studiengangsdaten wurden anschliessend mit Pandas vorab grob bereinigt. Danach wurden die drei Datenset aus dem Crawling in mehreren Schritten im Tableau Prep bereinigt, transformiert und verknüpft. Am Schluss konnte erfolgreich das Übersichtsdatenset in zwei Formaten (.csv und .tde) exportiert werden. Mit diesem Übersichtsdatenset konnten die ersten zwei Fragestellungen mit Hilfe von Tableau beantwortet werden. Für die dritte Fragestellung wurde ein weitere Tableau Prep Prozess erarbeitet. Das aus diesem Prozess entstandene Veränderungsdatenset umfasst nur noch Studiengangs-Einträge, welche für diese Zeitperiode als Zugänge oder Abgänge gelten und kennzeichnet die Einträge entsprechend. Mit diesem neu erarbeiteten Veränderungsdatenset konnte dann auch die dritte Fragestellung beantwortet werden. Zum Schluss wurde das Übersichtsdatenset in die Datenbank „Data Science Studiengangübersicht“ importiert.

Keywords

Webcrawler, Data Science, CIP, ETL, SQL Server, Tableau Prep, Tableau

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einleitung

Wir möchten aktuelle und nachfolgende Studierende dabei unterstützen ein Studium oder Austauschsemester im Bereich Data Science durchzuführen. Da momentan ein Boom im Bereich dieser Studienrichtung stattfindet und jährlich neue Programme entstehen, möchten wir in einer einfachen Übersicht das aktuelle Angebot von Bachelor und Master Studienprogrammen weltweit wiedergeben. Als zusätzliche Hilfestellung möchten wir noch die Ratingwertung der Unis einpflegen.

1.1 Aufgabenstellung

Ziel dieser Projektarbeit ist es, ein Übersichtsdatenset zu Studiengängen im Fachbereich „Data Science“ zu erstellen. Die dazu benötigten Daten werden aus dem Internet gecrawlt. Anschliessend sollen die Daten bereinigt und transformiert werden und als Übersichtsdatenset in einer SQL-Datenbank abgelegt werden. Zum Schluss soll mit diesem Dataset folgende drei konkrete Fragestellungen beantwortet werden.

1. Welches Masterstudienangebot für Data Science existiert in Europa?
2. Welche Universität in Asien, hat das beste internationale Ranking und bietet einen Studiengang in Data Science an?
3. Wie verändert sich das Studienangebot im Bereich Data Science in der Region Amerika im Verlauf der Zeit?

Die dritte Fragestellung kann zum Zeitpunkt der Abgabe der Projektarbeit wohlmöglich nicht beantwortet werden, da die dazu benötigten Daten erst nach Implementation des Crawlers periodisch (z.B. monatlich) erfasst werden können. Es soll aber zumindest aufgezeigt werden, wie der Prozess dafür aussieht. Im besten Fall lassen sich bereits Ergebnisse vorweisen.

1.2 Vorgehensweise

Um ein Übersichtsdatenset für die Studiengänge zu erstellen sind folgende Schritte geplant. Zuerst müssen die nötigen Daten aus dem Internet gecrawlt werden. In einem nächsten Schritt sollen zuerst die erarbeiteten Datensets einzeln mit dem Pandas Package bereinigt und transformiert werden, falls nötig. Anschliessend sollen die Datensets im Tableau Prep zu einem Datenset vereinigt werden und falls nötig nochmals

bereinigt und transformiert werden. Im letzten Schritt soll das erstellte Datenset in eine SQL-Datenbank importiert werden. Die Fragestellungen sollen mit Hilfe des erstellten Datensets und Tableau beantwortet werden.

1.3 Datenquellen und Daten

In diesem Abschnitt werden kurz die verschieden Datenquellen vorgestellt.

Für alle Data Science Studiengänge wurden folgende Seiten gecrawlt:

- Bachelor Studiengänge: https://www.bachelorstudies.com/Data-Science/
- Master Studiengänge: https://www.masterstudies.com/Data-Science/

Für die Hochschulrangliste wird folgende Seite gecrawlt:

- Ratings: http://www.webometrics.info/en/world

1.4 Verwendete Tools

In diesem Abschnitt werden kurz die verschiedenen Tools vorgestellt, die verwendet wurden.

1.4.1 Pycharm

Das Crawling und die erste Bearbeitung der Daten findet in PyCharm statt. PyCharm erlaubt es Python in einer nutzerfreundlichen Oberfläche anzuwenden. Es bietet auch intelligente Unterstützung, wie zum Beispiel Codevervollständigung oder Fehlerhervorhebung. Hier wird ein kurzer Überblick über die benutzten Packages und Module gegeben.

Pandas - Package

Pandas erlaubt es schnell und flexibel Dataframe Objekte mit integrierter Indexierung zu erstellen. Zudem erlaubt es das flexible Bearbeiten, Transformieren und Aggregieren von Daten. Auch bietet es die Möglichkeit Daten zu lesen und zu speichern.

csv - Package

Das csv-Package erlaubt es CSV-Dateien zu lesen und zu schreiben.

Requests - Modul

Dieses Modul erlaubt es http-Anfragen mittels Python zu senden. Die Anfrage gibt ein Antwortobjekt zurück, welche Antwortdaten beinhaltet. Darunter gehören Inhalt, Verschlüsselung, Status etc..

BeautifulSoup - Modul

Vom bs4-Package wurde das BeautifulSoup Modul verwendet. Das Modul erlaubt es gezielt einzelnen Elemente aus HTML oder XML-Code zu extrahieren.

Time – Package

Dieses Package umfasst eine Vielzahl von nützlichen Features. In diesem Fall wurde es nur zum Einbauen einer Wartezeit bei der Abfrage der Webseite verwendet.

1.4.2 Tableau Prep

Tableau Prep ist eine Applikation, welche ohne grosse Vorkenntnisse das Kombinieren, Formatieren und Aufbereiten von Daten für die Analyse erlaubt.

1.4.3 Tableau

Mit Tableau lassen sich mit wenigen Klicks Daten zusammenstellen und visualisieren. Veränderungen im Input werden direkt visualisiert. Auch lassen sich Dashboard erstellen.

1.4.4 SQL-Datenbank

Als Datenbank wurden Microsoft 2017 SQL Server verwendet. Diese Datenbank unterstützt Echtzeitanalysen.

1.4.5 Microsoft SQL Server Management Studio

Zur Verwaltung der Datenbanken wird die Microsoft SQL Server Management Studio Software verwendet. Sie erlaubt es in wenigen Schritten Datenbank zu erstellen und Daten zu importieren. Auch können in der Software Datenbankabrufe durchgeführt werden.

2. Prozessschritte zur Erstellung des Übersichtsdatensets

In diesem Abschnitt werden die einzelnen Schritte vom Crawling bis zum Transferieren des erstellten Übersichtsdatensets in die SQL-Datenbank beschrieben.

2.1 Collection und Extraction mit Python

Für die Collection und Extraktion der Daten wurden zwei Skripte benötigt. Das erste Skript wird für die beiden Studiengangs-Webseiten verwendet und das zweite für Ranglisten-Webseite. Vom Aufbau sind die beiden Skripte sehr ähnlich, unterscheiden sich aber beim Crawlerteil, bedingt durch die unterschiedliche Architektur der Webseiten. Zudem fehlt beim Skript für die Ranglisten-Webseite die Funktion „cleanse_data“, damit noch mehr Bereinigung- und Transformationssschritte im Tableau Prep gemacht werden können. In diesem Abschnitt werden die einzelnen Funktionen kurz beschrieben. Die vollständigen Skripte sind im Projektordner unter dem Ordner Python abgelegt.

Bibliothek vorbereiten

Die verwendeten Packages und Module werden im Kapitel 1.4.1 genauer vorgestellt.

Funktion send_request

Diese Funktion ruft den angegeben Pfad und seinen HTML-Code auf und gibt an ob die Anfrage erfolgreich war oder nicht.

Funktion collect_items

Ruft über die send_request Funktion den HTML Code des gewählten Pfads ab. Geht durch alle Seiten, die unsere Daten beinhaltet. Mit Hilfe des BeautifulSoup Modul werden die interessanten Daten des HTML-Code transformiert und in einer Liste “items” gespeichert. Ein Listeneintrag entspricht einer gecrawlten Seite.

Funktion extract_features

Extrahiert die einzelnen Datenfelder aus der “items”-Liste und speichert diese pro Eintrag (Schule) die in einer neuen Liste “content_element”. Alle “content elements” werden in einer Liste “content table” zusammengefasst. Zum Schluss wird die Liste noch in eine Pandas Dataframe umgewandelt und die Spalten beschriftet.

Funktion cleanse_data

Bei dieser Funktion werden die Daten der gecrawlten Data Science Hochschulen mit Hilfe vom Pandas Packages bereinigt. Als Bereinigungsschritte wurden Leerschläge und Zeilenumbrüche entfernt.

Funktion export_to_file

Mit dieser Funktion wird das bereinigte Dataframe als CSV-Datei exportiert.

Funktion main

Diese Funktion verknüpft alle vorhergegangen Funktionen zu einem Prozess, d.h. damit werden die Funktionen ausgeführt. Sie sammelt also zuerst alle Daten, extrahiert und transformiert dann die gewünschten Elemente in ein Dataframe. Beim Crawler Schools wird das Dataframe noch einem Bereinigungsschritt unterzogen vor dem Export. Zum Schluss wird das Dataframe als CSV-Datei exportiert.

2.2 Data Cleaning mit Pandas

Die Anwendung von Pandas und damit der erste Teil der Datenbereinigung erfolgt bereits in den Crawler Skripten (siehe Kapitel 2.1). Der Bereinigungs-Prozess wurde in der Funktion “cleanse_data” definiert. Dieser wurde nur bei den Studiengangs-Daten angewendet, damit mehr Bereinigungsschritte mit Tableau Prep übrigbleiben.

2.3 Data Cleaning, Integration und Transformation in Tableau Prep

Zur Verfügung stehen nach den drei Crawling-Prozessen drei CSV-Dateien: school_bachelor.csv, school_master.csv und school_ranking. Im nächsten Schritt werden diese mit Hilfe von Tableau Prep weiter bereinigt und transformiert. Damit bleibt am Schluss nur noch ein Datenset für die Beantwortung der Fragestellung übrig. Ziel ist es einen möglichst automatisierten Prozess im Tableau zu haben, damit bei neuen oder aktualisierten Daten, die Bearbeitung effizient stattfinden kann.

Einträge wurden, wenn möglich mit einer Logik z.B. einer Formel geändert, um systematische Probleme zu adressieren. Leider war es unumgänglich vereinzelt offensichtlich fehlerhaft Einträge zu ersetzen. Ein gutes Beispiel dafür ist: «24 years» statt «24 months». Es ist in diesem Fall schlichtweg unrealistisch, dass ein Studium 24 Jahre dauern soll. Hinzu kommt, dass beim Verknüpfungsschritt im grösseren Ausmass manuelle Korrekturen notwendig waren, da Tableau Prep kein partielles Matching erlaubt.

Im kommenden Abschnitt werden die einzelnen Prozessschritte kurz erklärt. Für Details sollte die Tableau Prep Datei “Übersichtdsdatenset_Prozess.tfl” konsultiert werden. In Abbildung 1-3 sind die einzelnen Schritte in einem Prozessschema visualisiert.

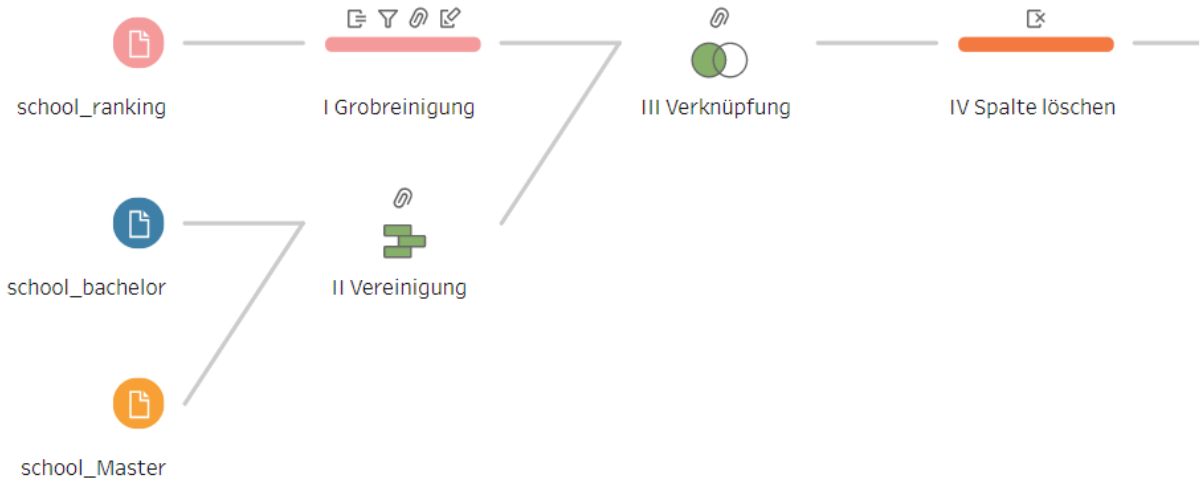


Abbildung 1 Prozessschema von Tableau Prep vom Import der Daten bis zu Schritt IV Spalten löschen

I Grobreinigung der “school_ranking” Daten

Die Bereinigung beschränkt sich hauptsächlich auf drei Schritte: Zahlen und Leerzeichen wurden aus den Namen, eine leere Zeile entfernt und der Spaltennamen angepasst. Heisst dieser nämlich gleich wie im zu vereinenden Datensatz, erkennt dies Tableau Prep automatisch.

II Vereinigung der Studiengang-Datensets

Die beiden CSV-Dateien mit den Studiengangs-Daten (school_bachelor.csv und school_master.csv) können unkompliziert kombiniert werden, da sie denselben Aufbau haben. Zusätzlich mussten einzelne Einträge manuell ersetzt werden. Dies wurde

schlicht aus dem Grund gemacht, um das Matching mit dem Ranglisten Datenset zu verbessern.

III Verknüpfung des vereinigten Studiengang-Datensets mit dem Ranglisten Datenset

In diesem Schritt wird das vereinigte Studiengang-Datenset mit dem Ranglisten-Datenset über die gemeinsame Spalte “Schulnamen” miteinander kombiniert. Da bei dieser Kombination viele übereinstimmende Einträge fehlerhafterweise nicht erkannt wurden und deswegen nicht verknüpft wurden, mussten beinahe 100 Einträge manuell korrigiert werden. Das Matching funktionierte teilweise nicht, weil die Schulnamen in einer anderen Sprache oder leicht abgeändert in einem der beiden Datensets vorgelegen haben. Leider ist im Tableau Prep keine partielles Matching möglich.

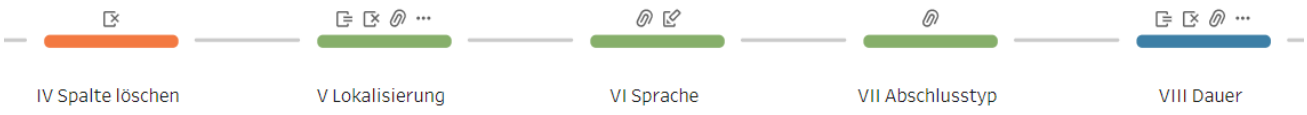


Abbildung 2 Prozessschema von Tableau Prep vom Schritt IV Spalten löschen zum Schritt VIII Dauer

IV Transformation - Spalte löschen

Nach dem Verknüpfen der Datensets bleiben zwei Schulspalten übrig. Es wurde die Schulspalte, die vom Ranglisten-Datenset abstammten, entfernt. Zusätzlich wurde table-Spalte entfernt, welche den Namen der Studiengang-Datensets beinhaltet.

V Transformation - Lokalisierung

Die Location-Spalte wurde in mehreren Operationen in die drei Spalten (Land, Stadt, Kontinent) transformiert. Die Operationen beinhalten das Berechnen von Felder, Gruppieren und Ersetzen und Einträgen und Umbenennen von Titelfelder. Zusätzlich wurde die Rolle der Spalten zu Land/Region geändert.

VI Transformation - Sprache

Die Languages-Spalte wurde in Sprache umbenannt. Zusätzlich wurde Einträge gruppiert und in einheitliche Einträge zusammengefasst. Wenn mehrere Sprachen angegeben wurden, wurden diese unter dem Begriff “Multiple” zusammengefasst.

VII Transformation - Abschlusstyp

Bei der Degree-Spalte wurden einige Einträge gruppiert und in einheitliche Einträge zusammengefasst. Zum Beispiel wurden Bachelor’s und BSC beide zu Bachelor zusammengefasst.

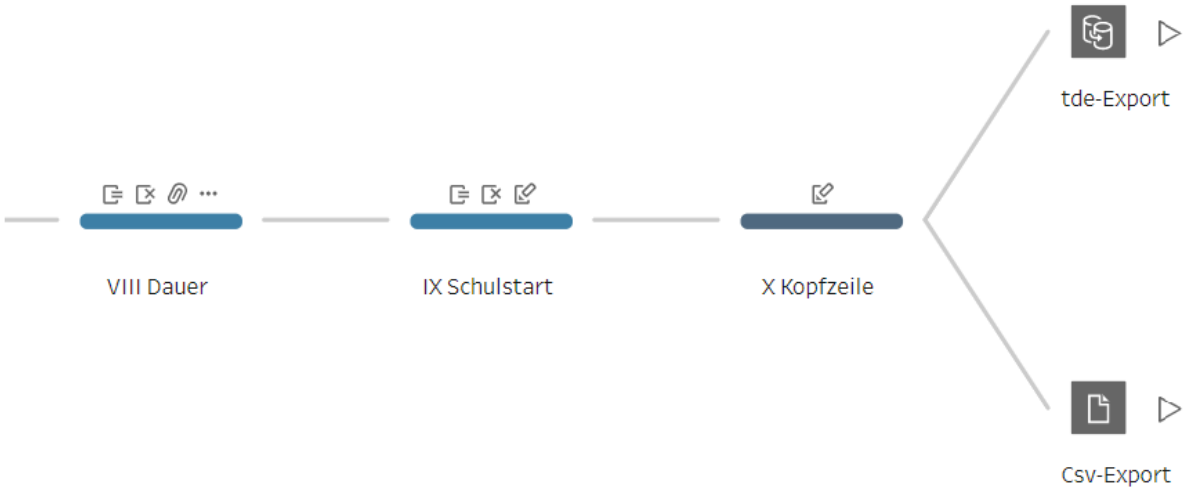


Abbildung 3 Prozessschema von Tableau Prep vom Schritt VIII Dauer bis zum Schritt Export

VIII Transformation - Dauer

Die Duration-Spalte wurden in mehreren Operationen in zwei Spalten (duration min und duration max) transformiert. Die Operationen beinhalten das Berechnen von Felder, Gruppieren und Ersetzen von Einträgen und Umbenennen von Titelfelder. Zusätzlich wurde die Rolle der Spalten zu Dezimalzahl geändert.

IX Transformation - Schulstart

Die Start-Spalte wurde in mehreren Operationen so bearbeitet, dass nur noch der Startmonat angegeben ist. Die Operationen beinhalten das Berechnen, Umbennen und Entfernen von Feldern.

X Transformation - Kopfzeile

Zum Schluss wurden noch diverse Kopfzeilen umbenannt. Ausserdem wurde die Reihenfolge der Spalten geändert.

Export der Datei

Das finale Datenset wurde in zwei Formaten exportiert. Einerseits als CSV-Datei um es in die SQL-Datenbank hochzuladen, aber auch um mögliche Visualisierung in R oder Python zu machen. Andererseits wurden die Daten als TDE-Datei exportiert für einen einfachen Import in Tableau.

2.4 Transfer der Resultate auf die Datenbank (SQL Server)

Als nächstes musste das erstellte Übersichtsdatenset noch so abgespeichert werden, damit es persistent abrufbar ist. Deshalb soll das erstellte Datenset langfristig in eine SQL-Datenbank importiert werden. Nach dem Tableau Prep Prozess steht das Übersichtsdatenset als CSV oder TDE-Datei zur Verfügung. Für einen einfach Import sollte die CSV-Datei verwendet werden. Im folgenden Abschnitt wird kurz beschrieben wie man eine personalisierte Datenbank erstellt und die Daten als CSV-Datei in diese Datenbank importiert.

Datenbank erstellen

Eine neue Datenbank kann erstellt werden, wenn man mit Rechtsklick auf Databases klickt und „New Database.“ auswählt (Abbildung 4A). Dann öffnet sich ein neues Fenster, in dem man den Namen und optional noch den Besitzer festlegen kann (Abbildung 4B). In unserem Falls heisst die Datenbank „Data Science Studiengangübersicht“.

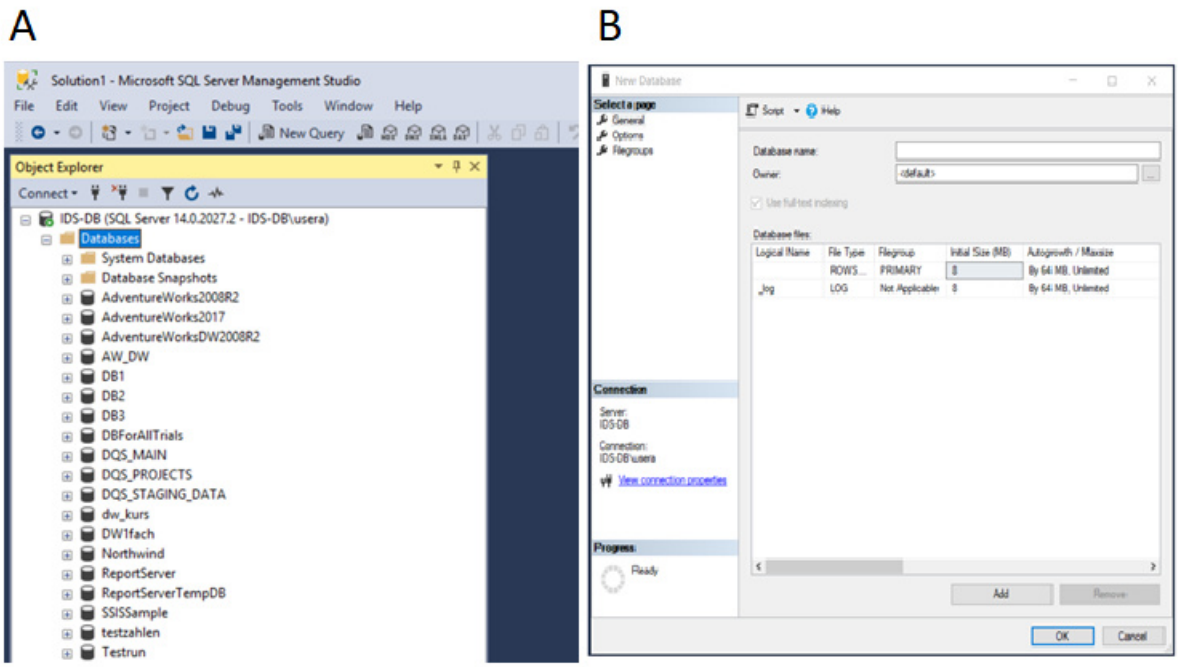


Abbildung 4 Explorer Fenster von Microsoft SQL Server Management Studio (A) und New Database-Fenster (B)

Datenset importieren

Nachdem nun die Datenbank erstellt ist, muss nur noch die CSV-Datei importiert werden. Dazu muss man die gewünschte Datenbank mit Rechtsklick auswählen, dann „Task“ auswählen und anschliessend „Import Flat File“ verwenden (Abbildung 5A). Im neuen Fenster kann man dann den Namen der Tabelle festlegen, sowie weitere Details. Es ist wichtig festzulegen welchen Datentyp die einzelnen Spalten haben und dass überall Nullwerte erlaubt sind. In unserem Fall erlauben wir allen Spalten mit dem Datentyp „Character“ bis zu 350 Wörter zu beinhalten, um sicher genug Zeichen zu Verfügung zu stellen, um alle Information zu erfassen. Nun muss man nur noch den Import beenden und die Daten sind persistent in der Datenbank abrufbar.

A B

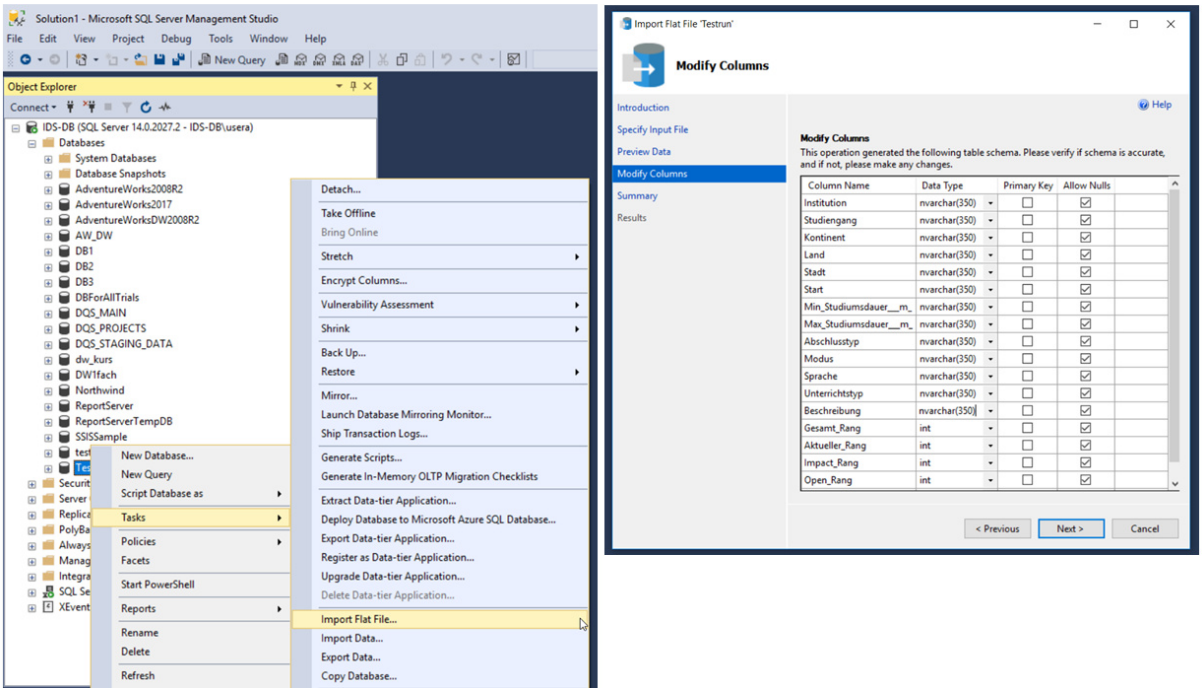
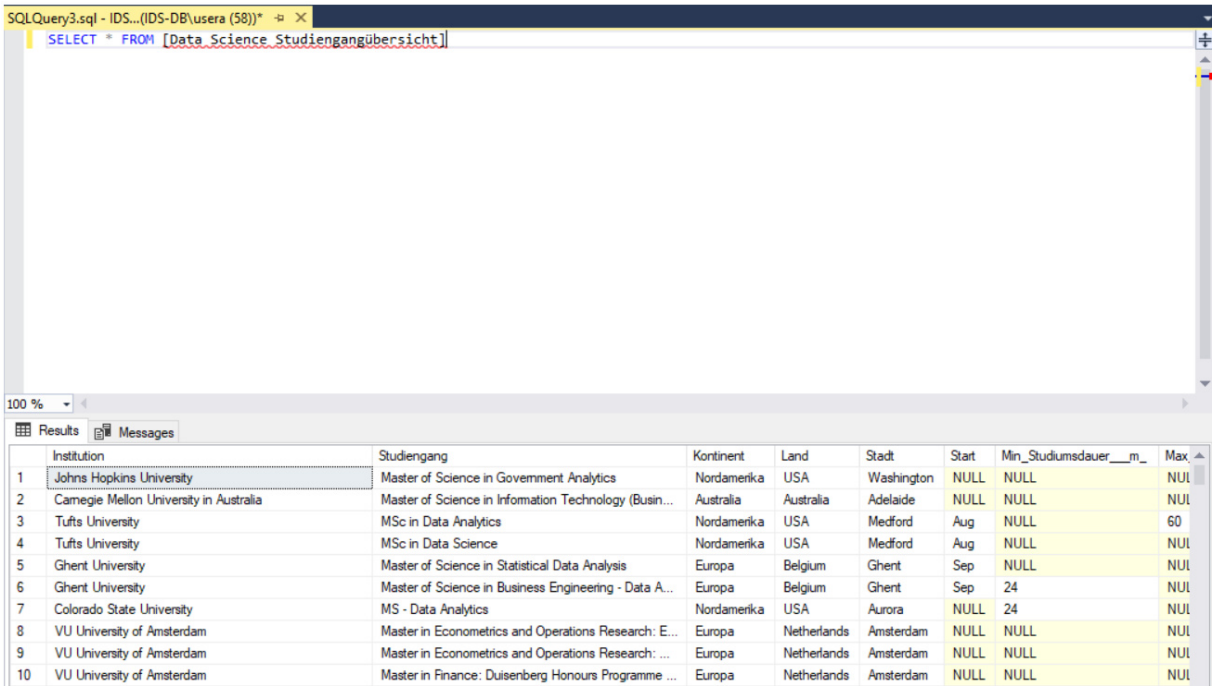


Abbildung 5 Explorer Fenster von Microsoft SQL Server Management Studio (A) und Import Flat File Fenster (B)

Kontrolle des Imports

Um zu kontrollieren, ob der Import funktioniert hat, kann man mit Rechtsklick auf die Datenbank klicken und „New Query“ auswählen. Dann muss man nur noch im Query Fenster „SELECT * FROM [Data Science Studiengangübersicht]“ eingeben (Abbildung 6) und die Anfrage ausführen. Nun sollten die Daten angezeigt werden.



	Institution	Studiengang	Kontinent	Land	Stadt	Start	Min_Studiumsdauer___m_	Max
1	Johns Hopkins University	Master of Science in Government Analytics	Nordamerika	USA	Washington	NULL	NULL	NUL
2	Carnegie Mellon University in Australia	Master of Science in Information Technology (Busin...	Australia	Australia	Adelaide	NULL	NULL	NUL
3	Tufts University	MSc in Data Analytics	Nordamerika	USA	Medford	Aug	NULL	60
4	Tufts University	MSc in Data Science	Nordamerika	USA	Medford	Aug	NULL	NUL
5	Ghent University	Master of Science in Statistical Data Analysis	Europa	Belgium	Ghent	Sep	NULL	NUL
6	Ghent University	Master of Science in Business Engineering - Data A...	Europa	Belgium	Ghent	Sep	24	NUL
7	Colorado State University	MS - Data Analytics	Nordamerika	USA	Aurora	NULL	24	NUL
8	VU University of Amsterdam	Master in Econometrics and Operations Research: E...	Europa	Netherlands	Amsterdam	NULL	NULL	NUL
9	VU University of Amsterdam	Master in Econometrics and Operations Research: ...	Europa	Netherlands	Amsterdam	NULL	NULL	NUL
10	VU University of Amsterdam	Master in Finance: Duisenberg Honours Programme ...	Europa	Netherlands	Amsterdam	NULL	NULL	NUL

Abbildung 6 SQl Query Fenster

2.5 Beantwortung der Fragestellungen

In diesem Abschnitt werden die drei Frage beantwortet. Die Fragestellung 1 und 2 wurden mit den gecrawlten Daten vom 20.03.2020 beantwortet.

Fragestellung 1: Welches Masterstudienangebot für Data Science existiert in Europa?

In Tableau wurden die Daten gefiltert, dass nur noch Hochschulen von Europa mit einem Master Programm angezeigt werden (Abbildung 7). 102 Hochschulen bieten momentan ein Masterprogramm an. Einige Hochschule bieten mehrere Programme im Bereich Data Science an. Die Hochschule Luzern fehlt leider in der Liste. Es wäre sinnvoll mit dem Webseitenbetreiber Kontakt aufzunehmen, dass auch unsere Schule in der Auflistung vorkommt.

Institution	Studiengang	Land	Start	Modus	Unterrichtstyp	Beschreibung	Gesamt-Rang
Università Commerciale Luigi Boc..	MSc in Data Science and Bus..	Italy	Null	Full-time	Campus	This MSc is designed..	1008
Aalto University Aaltoyliopisto	Master of Science in Comput..	Finland	Null	Full-time	Campus	Machine learning is ..	230
Arden Study Centre, Berlin	MSc Data Analytics and Info..	Germany	Apr	Full-time	Campus	Big Data is going to ..	Null
	MSc Data Analytics and Mar..	Germany	Apr	Full-time	Campus	Learn a range of ess..	Null
Arden University	MSc Data Analytics & Huma..	United Kingdom	Apr	Full-time	Online	The modern workpla..	9654
	MSc Data Analytics & Projec..	United Kingdom	Apr	Full-time	Online	The modern workpla..	9654
	MSc Data Analytics and Ente..	United Kingdom	Apr	Full-time	Online	Getting ahead of ch..	9654
Burgundy School of Business	MSc Data Science & Organis..	France	Sep	Full-time	Campus	The MSc Data Scienc..	6126
CESTE Business School	Master in Data Science	Spain	Null	Part-time	Campus	The Master aims to ..	Null
Charles University Faculty of Mat..	Master in Software and Dat..	Czech Republic	Null	Full-time	Campus	The study branch So..	Null
City University of London	MSc in Data Science	United Kingdom	Null	Full-time	Campus	Data science is an e..	496
Dalarna University College	Master in Data Science	Sweden	Sep	Full-time	Campus	Harness the power o..	2059

Abbildung 7 Auschnitt der erstellten Tableautabelle, welche die Beantwortung der Fragestellung 1 erlaubt

Fragestellung 2 Welche Universität in Asien, hat das beste internationale Ranking und bietet einen Studiengang in Data Science an?

In Tableau wurden die Daten gefiltert, dass nur noch Hochschulen von Asien mit einem Data Science Programm angezeigt werden (Abbildung 8). Zusätzlich wurden nur Hochschulen mit einem vorhandenem Hochschul-Ranking angezeigt werden. Es bleiben neun Hochschulen übrig, wobei die City University of Hong Kong das beste Ranking mit dem Gesamt-Rang von 186 vorweisen kann.

Institution	Studiengang	Land	Start	Modus	Unterric..	Beschreibung	Gesamt-Rang
Asian Institute of Management	Master of Science in Data Sc..	Philippines	Null	Full-time	Campus	A four-term, 14-mon..	5065
City University of Hong Kong	Bachelor of Engineering in C..	Hong Kong	Sep	Full-time	Campus	We aim to provide st..	186
	Bachelor of Engineering in D..	Hong Kong	Sep	Full-time	Campus	This major aims at e..	186
	Bachelor of Science in Data ..	Hong Kong	Sep	Full-time	Campus	This major is to prov..	186
Heriot-Watt University Dubai	BSc (Hons) in Computer Scie..	United Arab Emirates	Sep	Full-time	Campus	Our BSc Computer S..	510
	MSc in Data Science	United Arab Emirates	Sep	Full-time	Campus	You will learn how t..	510
Hong Kong Baptist University	MSc in Data Analytics and B..	Hong Kong	Sep	Full-time	Campus	The Master of Scien..	480
Middlesex University Dubai	MSc Data Science	United Arab Emirates	Sep	Full-time	Campus	The role of a data sci..	4686
Princess Nourah Bint Abdulrahma..	MSc in Computing (Data Ana..	Saudi Arabia	Sep	Full-time	Campus	This program opens ..	5667
Rochester Institute of Technology	MSc in Professional Studies:..	United Arab Emirates	Sep	Full-time	Campus	This program prepar..	421
Skolkovo Institute of Science and ..	Master of Science in Data Sc..	Russia	Sep	Full-time	Campus	Data scientists are g..	4559
Vietnam National University Hanoi	Bachelor in Business Data A..	Vietnam	Sep	Full-time	Campus	The Bachelor in Busi..	1133

Abbildung 8 Auschnitt der erstellten Tableautabelle, welche die Beantwortung der Fragestellung 2 erlaubt

Fragestellung 3 Wie verändert sich das Studienangebot im Bereich Data Science in der Region Amerika im Verlauf der Zeit?

Um die dritte Fragestellung zu beantworten ist es sinnvoll einen weiteren Tableau Prep Prozess zu erstellen. Bei diesem Prozess werden zwei Übersichtsdatensets von verschiedenen Erfassungszeitpunkten verwendet. Aus diesen zwei Datensets wird ein neues Datenset erstellt, welches nur Zugänge und Abgänge an Studiengängen beinhaltet. Das Schema des Tableau Prep Prozess ist in Abbildung 9 ersichtlich.

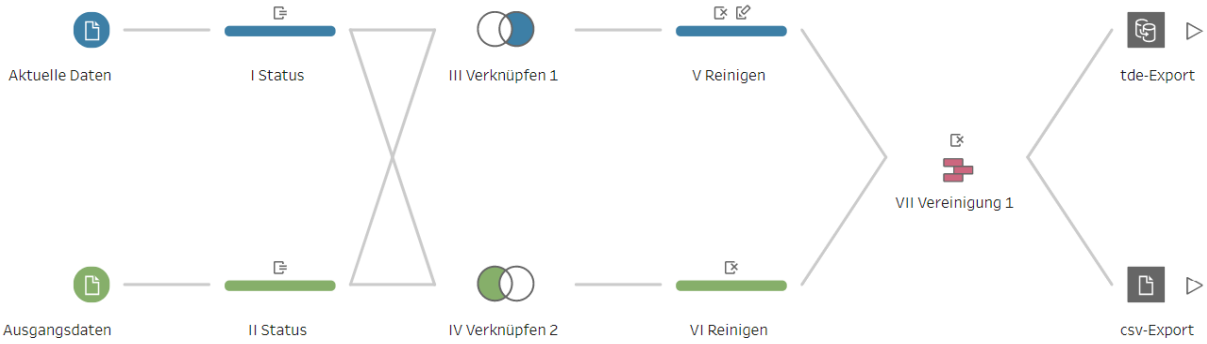


Abbildung 9 Prozessschema von Tableau Prep vom Import bis zum Export

Im folgenden Abschnitt wird der Prozess grob erläutert und das Resultat präsentiert. Den Prozess kann man im Detail in der Tableau Prep Datei „Veränderungsdaten-set_Prozess.tfl“ mitverfolgen.

Tableau Prep Prozess

Als erstes muss man zwei Übersichtsdatensets ins Tableau Prep importieren. In unserem Fall wurden diese am 20.03.2020 und 09.04.2020 erstellt. Nach dem Importieren wird bei beiden eine weitere Spalte hinzugefügt. Diese beinhaltet entweder das Wort Abgänge für die Ausgangsdaten und Zugänge für die aktuellen Daten. Nun werden die beiden Datensets verknüpft, wobei nur die Einträge behalten werden, welche nicht in beiden Datensets vorkommen. Es werden zwei Verknüpfungen durchgeführt. Eine für die Abgänge und eine für die Zugänge. Die beiden neuen Schnittmengen werden dann in einem weiteren Bereinigungsschritt für die Vereinigung vorbereitet. Bei diesem Bereinigungsschritt werden überflüssige Spalten gelöscht und Spaltentitel angepasst. Die beiden Schnittmengen können danach problemlos vereinigt werden. Im letzten Schritt wird das neu gewonnen Veränderungsdatenset als CSV und TDE-Datei exportiert. Diese kann nun in Tableau betrachtet werden.

Resultat

Das Studienangebot hat sich im Zeitabschnitt vom 20.03.2020 und 09.04.2020 für den Kontinent Nordamerika nur leicht verändert (Siehe Abbildung 10). Hinzugekommen ist

der Studiengang „MS in Health Data Analytics (On-Campus)“ der University of Louisville. Weggefallen ist hingegen der Studiengang “Master of Science in Computer Science Data Analytics» der Nova Southestern University.

Institution Status	Institution	Studiengang	Land	Start	Modus	Unterrichtstyp	Beschreibung	Gesamt-Rang
Abgänge	Nova Southeastern ..	Master of Science in ..	USA	May	Full-time	Campus	The Master of ..	724
Zugänge	University of Louisvi..	MS in Health Data A..	USA	Aug	Full-time	Campus	The Master of ..	353

Abbildung 10 Erstellte Tableautabelle, welche die Beantwortung der Fragestellung 3 erlaubt

Mit den beiden erstellten Tableau Prep Prozessen kann man nun jeden Monat mit den gecrawlten Daten ein neuen Übersichtsdatenset erstellen und die Veränderung des Angebots in einem Veränderungsdatenset festhalten. Der Vorteil der getrennten Prozesse ist, dass man die Veränderung des Studienangebot nun in einem selbst bestimmten Zeitfenster durchführen kann und dieses auch flexibel ändern kann.

3. Schlusswort

In der vorliegenden Projektarbeit wurde auf Basis von persönlichen Präferenzen, eine weltweite Übersicht für die Studiengangsrichtung “Data Science“ erarbeitet. Das Übersichtsdatenset liegt in mehreren Formaten zur einfachen Einsicht vor und wurde auch in die SQL-Datenbank „Data Science Studiengangübersicht“ importiert. Um dieses Übersichtsdatenset zu erstellen, wurde ein mehrstufiger Prozess durchgeführt. Der erste Schritt war die Webseiten, welche die Informationen beinhalteten, mit Hilfe von BeautifulSoup zu crawlen, die essenziellen Daten zu extrahieren und mit Pandas vorab zu bereinigen. Danach mussten die verschiedenen Datensets in mehreren Schritten im Tableau Prep zu einem Übersichtsdatenset vereint werden. Diese Schritte beinhaltete Transformation und Bereinigung von Daten. Mit diesem Übersichtsdatenset und Tableau konnte erfolgreich die ersten zwei Fragestellungen beantwortet werden. Für die dritte Fragestellung wurde, ein weiterer Tableau Prep Prozess erstellt, welcher aus dem Übersichtsdatenset von zwei verschiedenen Zeitpunkten ein Veränderungsdatenset erstellt. Dieses Datenset beinhaltet nur noch Einträge von neuen, aber auch verschwundenen Studiengängen der Fachrichtung „Data Science“ weltweit. Mit diesem Datenset lässt sich dann die dritte Fragestellung für einen gewählten Zeitraum beantworten.

3.1 Diskussion

Die erarbeiten Datenset erlauben es die Fragestellungen vollumfänglich zu beantworten. Die Crawler sollten mit ihrem jetzigen Aufbau langfristig funktionieren. Nur die gecrawlten Daten der Studiengänge wurden mit Pandas gereinigt. Es wäre sinnvoll diese bei einer Weiterverfolgung des Projekts bei allen gecrawlten Daten durchzuführen. Auch macht es Sinn die Bereinigungsschritte in Pandas auszubauen, um den Tableau Prep Prozess zu entschlacken. Der Tableau Prep funktioniert soweit wie gewünscht. Problem könnte aber in Zukunft entstehen, wenn neue Hochschulen Data Science Studiengänge einführen. Das Problem liegt beim Verknüpfungsschritt im Tableau Prep. Die Datensets der Studiengänge und der Rankings werden über den Hochschulnamen verknüpft. Die Namen derselben Hochschulen unterscheiden sich leider in nicht reproduzierbarer Weise. Zum Beispiel ist der Namen in einer anderen Sprache

hinterlegt. Dieses Problem kann man mit den verwendeten Webseiten höchstens minimieren, durch Kontrolle und optimieren der Reinigungsschritte und Anpassung der Namensvariationen. Für eine 100 % Matching müsste man alle Datensets mittels einer klar zuweisbaren ID oder örtlichen Adresse zur Verfügung haben.

3.2 Reflexion

Es war erstaunlich wie schnell man einen simplen Crawler erstellen kann, der funktioniert. Sobald aber mehr Details gewünscht werden oder die Webseitenarchitektur ändert wird es schnell sehr herausfordernd. Aufgefallen ist auch, dass Tableau Prep einen schnellen Einstieg in die Datenbereinigung erlaubt und schnell tolle Ergebnisse möglich sind. Trotzdem sind für Spezialisten mit Pandas bessere Ergebnisse möglich, da mehr Möglichkeiten zur Verfügung stehen. Auch ist der Prozess als Experte im Python Skript einfacher nachzuvollziehen.

3.3 Ausblick

Der nächste Schritt in diesem Projekt wäre klar die Verknüpfung der verschiedenen Datensets zu optimieren. Des Weiteren könnten das erstellte Übersichtsdatenset zur Entwicklung einer Applikation verwendet werden. Die Applikation könnte Jugendlichen bei der Studiums Planung helfen. Es könnte auch HSLU Studenten helfen eine geeignete Hochschule für ein Austauschsemester zu evaluieren. Zusätzlich könnte es der Schulleitung helfen geeignete Partneruniversitäten für einen Zusammenarbeit zu evaluieren.