Bericht Gipfelstürmer

Für dieses Projekt wurden Daten von www.hikr.org gescraped und analysiert. Dies wurde mit Scrapy und Spark gemacht. Dieser Bericht beschreibt das Vorgehen hierfür, die Qualität der Daten und allfällige Hinweise zum Crawling dieser Seite.

Datenqualität

Die Datenqualität auf hikr.org ist im Allgemeinen als gut zu bewerten, weist jedoch in bestimmten Bereichen typische Mängel nutzergenerierter Inhalte auf. Die Analyse der Datenqualität beschränkt sich auf die für die Arbeit genutzten Attribute. Diese sind: **Name der Tour, Region, Datum, Aufstieg in Metern, Abstieg in Metern und die Gipfel**. Die Vollständigkeit der Daten ist in der nächsten Grafik visuell anhand der verschiedenen Attribute dargestellt.

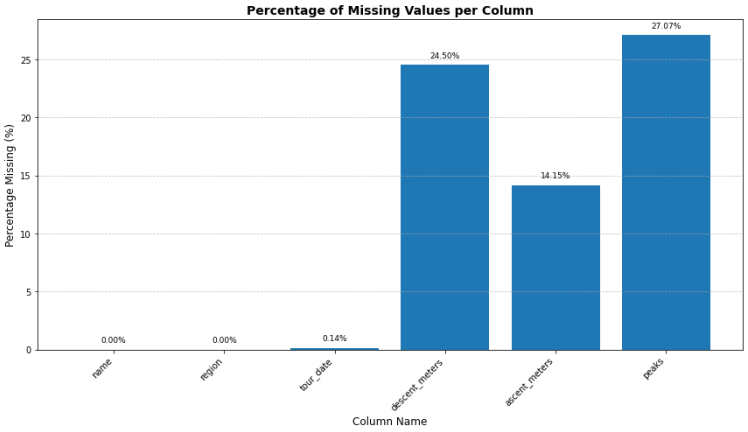

Abbildung 1: Fehlende Werte pro Attribut in %

Die Grafik zeigt, dass Basisinformationen wie der Name der Tour und die Region lückenlos vorhanden sind. Auch das Tourdatum ist grösstenteils erfasst. Deutlichere Datenlücken zeigen sich bei den Höhenmetern für Auf- und Abstieg. Dies könnte daran liegen, dass Nutzer diese Angaben nicht immer konsequent eintragen oder manche Touren, wie flache Wanderungen, keine signifikanten Höhenunterschiede aufweisen. Ebenso listet mehr als ein Viertel der Berichte keine spezifischen Gipfel auf, was plausibel ist, da nicht jede Tour zwangsläufig einen Gipfel zum Ziel hat, wie beispielsweise bei Talwanderungen oder Hüttentouren.

Bezüglich der Korrektheit und Konsistenz der Daten ist festzuhalten, dass Datumsangaben als Freitext mit deutschen Monatsnamen und manchmal inkonsistenten Jahresformaten (z.B. «218» für 2018 oder «9» für 2009) vorliegen. Deshalb ist hier eine robuste Parsing-Logik unerlässlich. In dieser Arbeit wurde das Datum dann in ein einheitliches Format ("dd.mm.yyyy") überführt. Dies gelang mit dem robusten Parsing für alle nicht-leeren Datumsangaben. Die Angaben zu Auf- und Abstieg sind zwar numerisch, enthalten aber häufig die Einheit "m". Diese muss entfernt werden, um mit den Zahlen arbeiten zu können. Es konnten alle nicht leeren Angaben zum numerischen Typ konvertiert werden. Allerdings deuten deskriptive Statistiken auf potenzielle Eingabefehler oder Ausreisser hin, wie extreme Maximalwerte (z.B. Aufstieg > 140’000m) und negative Werte (z.B. Abstieg -2200m). Diese müssen kritisch betrachtet werden, da sich diese Werte vielleicht erklären lassen (z.B. eine Mehrtägige Tour, welche zu vielen Höhenmetern führt) oder es sich um Fehler handeln kann. Die Regionsangaben, dargestellt als hierarchischer Pfad (Breadcrumbs), sind konsistent, erfordern jedoch ein sauberes Extrahieren und Zusammenfügen der einzelnen Teile.

Crawling

Beim Crawling von hikr.org sollten verschiedene Aspekte berücksichtigt werden, um eine erfolgreiche Datenextraktion zu gewährleisten. Die Webseite basiert grösstenteils auf purem HTML. Die Struktur der Webseite zeigt, dass relevante Daten in der Regel in Tabellen (<table>)

Benaja Kuhn

und spezifischen HTML-Elementen mit Klassen wie «fiche_rando_b» (für die Feldbezeichnung) und «fiche_rando» (für den Wert) zu finden sind. XPath- und CSS-Selektoren erweisen sich hier als gut geeignet. Da es sich um nutzergenerierte Inhalte handelt, muss ein Crawler robust gegenüber kleineren Abweichungen oder fehlenden Feldern konzipiert sein. Deshalb lohnt es sich z.B. Whitespaces zu entfernen und die Felder bei fehlenden Elementen auf «None» zu setzen.

Datenaggregation

Im Rahmen der Analyse wurde eine Aggregation durchgeführt, um die durchschnittlichen Auf- und Abstiegsmeter sowie die Anzahl der Touren pro Jahreszeit zu ermitteln. Die Jahreszeiten wurden basierend auf dem Monat des Tourdatums definiert: Frühling (März, April, Mai), Sommer (Juni, Juli, August), Herbst (September, Oktober, November) und Winter (Dezember, Januar, Februar). Die Resultate der Aggregation sind in den folgenden zwei Abbildungen gezeigt.

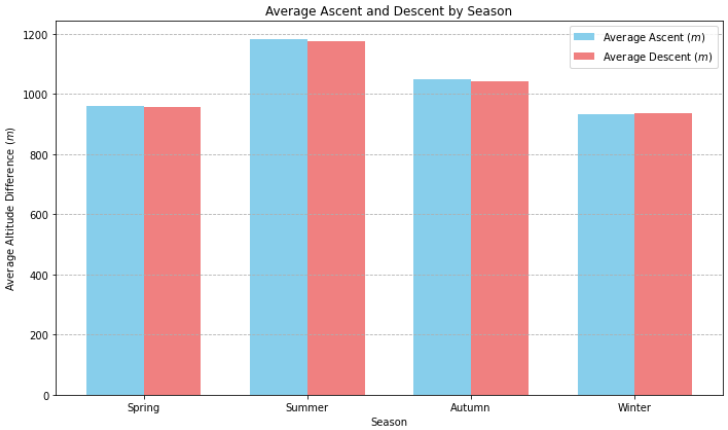

Abbildung 2: Durchschnittliche Anzahl Höhenmeter beim Auf- und Abstieg nach Jahreszeit

Diese Aggregation illustriert, dass die meisten Touren im Sommer unternommen werden, gefolgt vom Herbst, mit den wenigsten Touren im Winter. Die Anzahl der Höhenmeter im Auf- und Abstieg ist im Durchschnitt immer etwa gleich, wobei der Abstieg jeweils leicht kleiner ist. Dies könnte aber auf die höhere Anzahl fehlender Abstiegswerte zurückgeführt werden. Die Touren im Sommer weisen nebst der hohen Tourenzahl auch meisten durchschnittlichen Höhenmeter auf. Im Gegensatz dazu weisen Touren im Winter im Schnitt die geringsten Aufstiegsmeter auf. Solche saisonalen Aggregationen können wertvolle Einblicke in das Wanderverhalten liefern und bei der Planung zukünftiger Aktivitäten hilfreich sein.

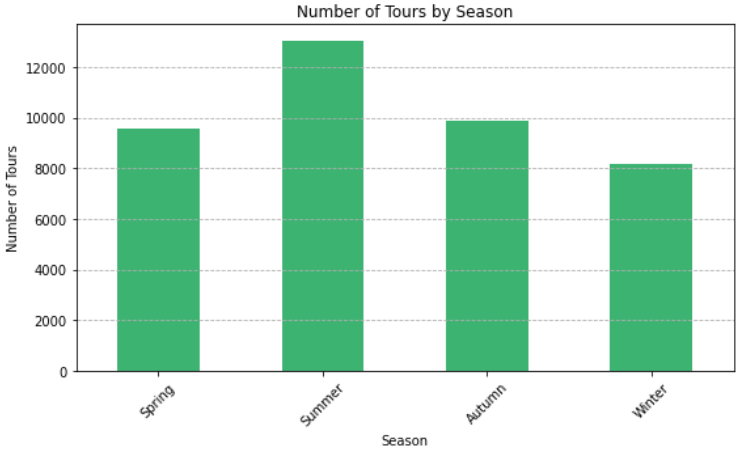

Abbildung 3: Anzahl Touren nach Jahreszeit

Fazit

Hikr.org stellt eine reichhaltige Datenquelle für Wanderberichte dar. Die Datenqualität ist für viele Attribute gut, erfordert aber, wie bei nutzergenerierten Inhalten üblich, eine sorgfältige Bereinigung und Validierung. Insbesondere bei Feldern wie Höhenmetern und Gipfelangaben ist mit Datenlücken zu rechnen. Zukünftige Entwickler sollten robuste Extraktionsmethoden verwenden und auf mögliche Inkonsistenzen oder Fehler in den Daten vorbereitet sein.

Benaja Kuhn