Datenanalyse mit Scrapy und Spark

Feature-Extraktion

Aus den Berichten wurden folgende Features extrahiert:

- Name
- Land
- Region (in der Schweiz: Kanton)
- Schwierigkeitsgrade: Klettern / Eisklettern / Klettersteig / Wandern / Hochtouren / Ski / Mountainbike
- Zeitbedarf
- Aufstieg
- Abstieg
- Wegpunkte
- Gipfel

Diese Features wurden aus dem HTML der Berichte gelesen. Für alle Elemente wurden Abstände, Zeilenumbrüche und Tabs entfernt. Danach wurden für einzelne Features individuelle Bereinigungen betrieben:

- Zeitbedarf: Damit die Werte der verschiedenen Touren miteinander verglichen werden können, wurden verschiedene Zeit-Formate zu Minuten umgerechnet.
- Aufstieg / Abstieg: Die Einheit wurde entfernt und der Wert in eine Ganzzahl konvertiert.
- Gipfel: Die Gipfel sind in verschiedenen Formaten in den Berichten enthalten. Wo möglich, wurde aus dem Text ein Objekt mit Namen, Höhe und ID generiert.

Auswertung mit Spark

Um die Rangliste der meistbesuchten Gipfel zu erstellen, wurden die vorhandenen Berichte nach den folgenden Kriterien gefiltert:

- Aufstieg grösser als 500 Meter
- Abstieg grösser als 500 Meter
- Zeitbedarf grösser als 120 Minuten
- Wanderung ist in der Schweiz

Anschliessend wurden die zehn Gipfel bestimmt, welche am häufigsten in den Berichten vorkommen. Ausgegeben werden genau zehn Gipfel. Haben zwei Gipfel gleich viele Referenzen in den Berichten, wird alphabetisch nach dem Namen sortiert. Aus dieser Abfrage resultiert folgende Tabelle:

Gipfel	**Anzahl Berichte**
Monte Generoso / Calvagione	62
Säntis	58
Schäfler	48
Girenspitz	42
Mutschen	40
Altmann	39
Lütispitz	39
Monte Gradiccioli	39
Monte Lema	39
Grosser Mythen	38

Datenanalyse

Aggregation

Aus den Berichten wurde für jedes Land der durchschnittliche Aufstieg, durchschnittliche Abstieg und der durchschnittliche Zeitbedarf berechnet. Das Resultat der Aggregation ist im Python Notebook sichtbar.

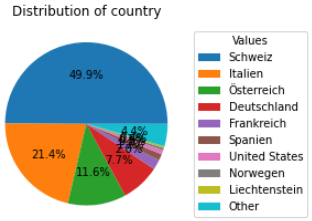
Bewertung der Datenqualität

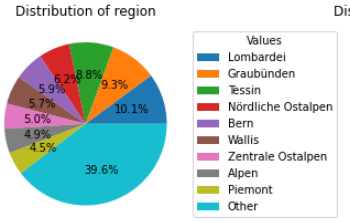
Bei der Analyse der vorhandenen Daten konnte festgestellt werden, dass ein erheblicher Teil der Berichte nicht vollständig ist, beziehungsweise keine Werte für die extrahierten Attribute hat. Für die Attribute der Schwierigkeitsgrade ist es erklärbar, wieso nicht alle Berichte einen Wert haben. So haben nicht alle Touren alle Elemente, nicht jede Tour bspw. hat einen Kletter-Abschnitt.

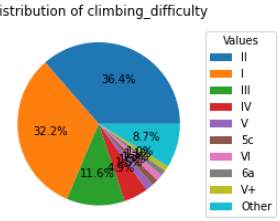
Für andere Attribute wie Aufstieg, Abstieg und Zeitbedarf sind die fehlenden Werte aber nicht plausibel zu erklären. Weil der Anteil der fehlenden Werte bei diesen drei Attributen 14.6, 24.7 und 26.6 Prozent beträgt, ist auch die berechnete Aggregation mit starker Vorsicht zu geniessen.

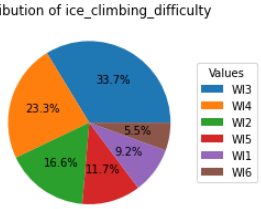
Aus der Verteilung der verschiedenen Attribut-Werten können weitere Aussagen zur Datenqualität gemacht werden:

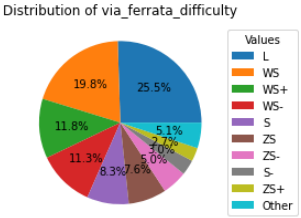
- Fast die Hälfte der Touren ist in der Schweiz. Die Datenlage für Touren in der Schweiz ist folglich grösser als in anderen Ländern. Somit können die errechneten Werte der Aggregation nicht als Referenzwerte für Touren in diesem Land verwendet werden, weil die Daten nicht unbedingt repräsentativ sind.
- Je höher der Schwierigkeitslevel, desto weniger Berichte gibt es für diese Tour. Es ist anzunehmen, dass es generell weniger schwierige Touren gibt, und auch weniger Personen eine schwierige Tour bestreiten, weswegen diese Verteilung nachvollziehbar ist.

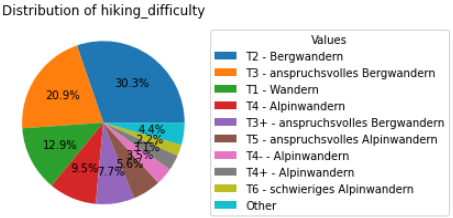


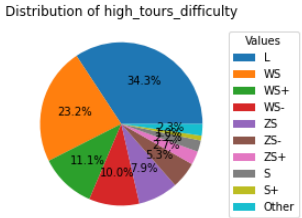


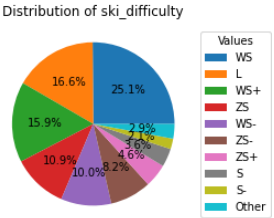


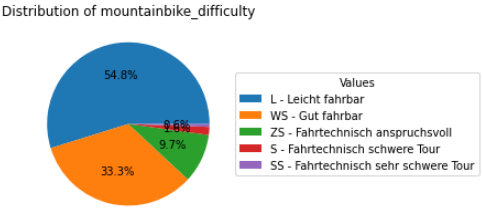












Bewertung der Korrektheit der Daten

Es wurden verschiedene Checks zu der Korrektheit der Daten durchgeführt:

- Höhe der Gipfel: Kein Gipfel ist höher als der höchste Berg der Erde (Mount Everest). Zudem haben die Gipfel keine negative Höhe.
- Übereinstimmung der Regionen und Länder: Für alle Touren in einer Schweizer Region wurde überprüft, ob auch das Land korrekt gesetzt ist (Spezialfall: Jura kann auch in Frankreich sein). Diese Überprüfung wurde vorerst nur für Schweizer Regionen umgesetzt, könnte aber auch für die restlichen Länder implementiert werden.
- Überprüfung der Schwierigkeitsgrade: Es wurde überprüft, ob die Schwierigkeitsgrade der Touren mit den Kategorien von SAC (Schweizer Alpen-Club) oder UIAA (Union Internationale des Associations d'Alpinisme) übereinstimmen. Dabei ist aufgefallen, dass die Schwierigkeitsgrade für Klettersteige nicht nach den SAC- oder UIAA-Normen bewertet werden, wie das bei beinahe alle anderen Touren gemacht wird.
- Überprüfung Aufstieg und Abstieg: Aufstieg und Abstieg sollen keine negativen Werte haben. Bei diesem Check ist aufgefallen, dass gewisse Touren negative Werte für den Abstieg haben. Der Parsing-Teil des Notebooks wurde daraufhin so angepasst, dass jeweils der Betrag des Auf- bzw. Abstieges verwendet wird, damit alle Werte positiv sind.
- Überprüfung benötigte Zeit: Keine Werte sind negativ