Assignment 3: Gipfelstürmer

Idee

Sie planen eine Wanderung für diesen Sommer. Im speziellen möchten Sie verschiedene Schweizer Gipfel besuchen und konsultieren dafür die Wanderberichte der Seite www.hikr.org um die beliebtesten Gipfel zu finden. Ebenso wollen Sie die weiteren Daten auf der Seite nutzen, um die Suche nach dem nächsten Wanderziel weiter einzugrenzen.

Aufgabenstellung

Teil 1: Feature-Extraktion mit Scrapy

Als Datenquelle haben Sie einen Abzug der Webseite www.hikr.org erhalten, auf welcher viele Wanderberichte verfasst sind. Sie sollen aus diesen Berichten mithilfe von Scrapy Attribute extrahieren, anhand deren Sie geeignete Gipfel auswählen. Zum Beispiel können Sie den Tourtyp des Berichtes auslesen, um Kletterberichte auszuschliessen. Extrahieren Sie mithilfe von Scrapy **mindestens 4 Attribute**, die Sie im zweiten Schritt dann zur Filterung verwenden können. Die Auswahl der Attribute ist Ihnen überlassen, soll aber zielführend sein. Für diesen ersten Teil werden Sie nur mit einer kleinen Teilmenge der Berichte arbeiten.

Im Jupyter Notebook soll klar verständlich werden, welche Features Sie gewählt haben und was für Datenbereinigungen notwendig waren.

Teil 2: Auswertung mit Spark

Im zweiten Schritt geht es darum, die Extraktion auf mehr Berichte anzuwenden. Eine sequenzielle Ausführung ist nicht geeignet aufgrund des hohen Rechenaufwandes. Kombinieren Sie Ihre Extraktionslösung daher mit Spark, um die Extraktion parallel für mehrere Berichte auszuführen.

Rangliste

Da Sie nun die Daten bereits in Spark haben, bietet sich eine weitere Auswertung mithilfe von Spark an. Erstellen Sie eine Rangliste der Gipfel. Filtern Sie dazu die Gipfel anhand der im Teil 1 extrahierten Attribute und sortieren Sie die Gipfel nach Anzahl Tourenberichten. Achten Sie auch darauf, dass die einzelnen Berichte 0 bis n Gipfel enthalten können. Erstellen Sie daraus eine Liste mit den 10 am stärksten besuchten Gipfeln, welche die von Ihnen definierten Kriterien erfüllen. Definieren Sie zudem, wie die die Rangliste sortiert wird, wenn 2 Gipfel gleich häufig besucht wurden.

Teil 3: Datenanalyse

Neben der Rangliste sollen Sie auch die Gesamtheit der Daten genauer analysieren. Analysieren Sie zum einen die Qualität der Daten, welche Sie gesammelt haben und bewerten Sie diese kritisch. Als weitere Analyse sollen Sie mithilfe von Spark eine Aggregation berechnen, bei welcher Sie ein Feature zum Beispiel räumlich oder zeitlich aggregieren. Die Aggregation soll dabei mindestens **4 Werte** haben. Die globale Durchschnittslänge aller Touren wäre somit nicht genügend, aber zum Beispiel die Durchschnittslänge pro Jahreszeit würde die Bedingungen erfüllen.

Bericht

Erstellen Sie einen **kurzen Bericht** auf Deutsch oder Englisch (ca. 1-2 Seiten A4) welcher sich an zukünftige Entwickler richtet, welche ebenfalls die Seite hikr.org scrapen möchten, allenfalls aber mit anderen Libraries und anderen Auswertungen. Dokumentieren Sie die generelle Qualität der Daten, aber auch worauf beim Crawling dieser Seite generell geachtet werden muss und wo Probleme,

mögliche Verzerrungen oder Datenlücken bestehen können. Beschreiben Sie auch kurz Ihre Aggregation, welche Sie im Rahmen dieser Analyse gemacht haben.

Hinweis: Der Bericht braucht NICHT die Resultate aus Teil 1 oder 2 zu wiederholen, es reicht, wenn diese im Jupyter Notebook sind.

Abgabe

Packen Sie Ihre Daten in ein .zip File mit dem Namen «**dawr_3_vorname_nachname.zip**». Dieses zip File soll enthalten: Ein Pdf mit dem Bericht (ca. 2 Seiten), ein Jupyter Notebook mit dem Code sowie benötigte Files mit Daten falls nötig. Maximum 5 MB, wenn die Daten grösser sind, können Sie auch einen Link im zip oder im Notebook (NICHT im Mail) hinterlegen. Das Jupyter Notebook muss vollständig ausführbar sein. Verwenden Sie keine absoluten Pfade im Code.

Senden Sie das .zip bis am **8. Juni 2025, 23:59:59** an lucas.broennimann@fhnw.ch

Code, welcher nicht von Ihnen selbst ist, muss entsprechend markiert sein. Plagiate werden mit der Note 1 bewertet. Nichteinhalten der Termine und Abgabebedingungen wird mit mindestens einer Note Abzug bestraft.