Big Data Projekt mit Yelp Datensatz

Projektmitglieder:

Nour-Eddine Kzaiber

Einführung:

Dieser Bericht präsentiert die Ergebnisse und Analysen eines Big-Data-Projekts, das sich auf den Yelp-Datensatz konzentriert. Ziel des Projekts ist es, mithilfe fortschrittlicher Datenanalyse Techniken wertvolle Einblicke in Unternehmen, Benutzerbewertungen und zugehörige Daten zu gewinnen. Durch die Analyse des Yelp-Datensatzes wollten wir nützliche Informationen extrahieren und umsetzbare Empfehlungen für Unternehmen bereitstellen.

Datensätze Übersicht:

Yelp ist ein internetbasiertes Unternehmen, das Bewertungen für Millionen von Restaurants und anderen Unternehmen enthält, die von Yelp-Nutzern bereitgestellt werden. Der Yelp-Datensatz umfasst verschiedene Daten zu Unternehmen, Benutzerbewertungen, Check-ins und Tipps. Es bietet eine umfassende Ressource zum Verständnis von Kunden Stimmungen, Geschäftsmerkmalen und Benutzer Präferenzen. Der Datensatz umfasst Informationen wie Unternehmenskategorien, Bewertungen, Benutzer Demografie, Bewertungstexte und mehr. Der Yelp-Datensatz, an dem dieses Projekt beteiligt ist, enthält über 5 Millionen Bewertungen von 1.621.950 Benutzern für etwa 115.527 Unternehmen und konzentriert sich auf Großbritannien, Deutschland, Kanada und in den USA von Februar 2005 bis Januar 2022.

Datensatz:

Quelle: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset

Projektbeschreibung & Ziele:

Die Idee besteht darin, die 5 Datendateien – Unternehmen, Benutzer, Bewertung, Tip & Check-in zu verwenden, um paar Fragen zu beantworten.

Wir möchten die nächsten Fragen beantworten:

- In welchem Jahr hatte Yelp die maximale Nutzerzahl?
- Prognose der Benutzer, die Yelp in den kommenden Jahren beitreten werden?
- Wer sind die aktivsten Benutzer auf Yelp?
- Best Bewertete State?

- Welche Benutzerbewertung ist basierend auf den Stimmen am beliebtesten? optional

- Welche State hat die höchste Anzahl geschlossener Geschäfte?

- Stimmungsanalyse zu den Top 10 Lebensmittelketten in den USA und Deutschland, basierend auf Nutzerbewertungen? optional

- Die besten geeigneten Restaurants für top Benutzer? optional

- Die 50 an den häufigsten verwendeten Worten in den Bewertungen?

Zeitplan zur Umsetzung:

Bitte werfen Sie einen Blick auf die Excel-Tabelle, um den Zeitplan für die Umsetzung des Projekts einzusehen.

	Aufgaben	**Status**	**Deadline**
Daten vorbereiten	Daten einlesen und kennenlernen		Montag
	Einfache Datenvisualisierung (z.B. Überprüfung auf Outlier)		Dienstag
	Preprocessing/ Datenbereinigung (Vorgehen beschreiben: z.B. Leerzellen entfernen)		Dienstag
	gewünschte Grafiken definieren		Dienstag
	Ggf. Machine Learning Algorithmen		Dienstag
Machine Learning Algorithmen	Machine Learning Algorithmen wie Regressionen, Mapreduce durchführen		Mittwoch
Dashboard	Layout designen		Dienstag
	Erstellen des Dashboards mit Platzhaltern		Mittwoch
	Einbau der Plots und Callback-Funktionen		Donnerstag
Präsentation	Code erklären		Freitag

Tools anzuwenden:

Python
Hadoop
Pig
Hive
Spark
Streamlit / Plotly Dash