Projektvorschlag: Big Data-Analyse zur Nutzung von sozialen Medien und emotionalem Wohlbefinden

Projekttitel: Analyse der Auswirkungen der Nutzung sozialer Medien auf das emotionale Wohlbefinden mithilfe von Big Data-Technologien

Vorbereitet von:

• Bratati Chakraborti

Betreut von:

• Prof. Axel Wemmel

**Ziel:**

Das Hauptziel dieses Projekts ist es, eine umfassende Datenanalyse unter Verwendung einer Reihe von Big Data-Technologien auf einem lokalen Rechner durchzuführen.

Das Hauptziel dieses Projekts ist es, die Nutzungsmuster verschiedener sozialer Medienplattformen (Facebook, LinkedIn, Snapchat, Twitter usw.) von Männern und Frauen zu analysieren und ihre Auswirkungen auf verschiedene Aspekte der psychischen Gesundheit wie Glück, Langeweile, Traurigkeit und Angst zu untersuchen. Das Projekt wird eine Reihe von Big Data-Technologien auf einem lokalen Rechner nutzen, um die Daten zu verarbeiten, zu bereinigen, zu verknüpfen, zu analysieren und zu visualisieren, um aussagekräftige Erkenntnisse zu gewinnen.

Datenquelle: Mehrere CSV-Dateien von Kaggle (Nutzung sozialer Medien und emotionales Wohlbefinden (kaggle.com))

**Werkzeuge und Technologien:**

Hadoop:

o HDFS (Hadoop Distributed File System): Zum Speichern von Daten auf einem lokalen Rechner über einen simulierten Cluster.

o MapReduce: Verwendet für die Verarbeitung großer Datensätze auf einem lokalen Rechner über WSL (Windows Subsystem für Linux) mit Ubuntu.

Apache Spark:

o PySpark: Python API für Spark zur Durchführung von Datenverarbeitung und -analyse in Jupyter Lab.

o Spark SQL: Für die strukturierte Datenverarbeitung und Ausführung von SQL-Abfragen.

o MLlib: Spark's Bibliothek für maschinelles Lernen zur Implementierung von Regressionsanalysen.

Pandas:

o Ein leistungsstarkes Python-Datenanalysetoolkit, das zur Bereinigung und Transformation der Daten verwendet wird.

Jupyter Lab:

o Eine interaktive Entwicklungsumgebung für Datenanalyse und -visualisierung.

o Verwendet für das Schreiben und Ausführen von Spark-Code sowie zum Erstellen von Notebooks, die Live-Code, Gleichungen, Visualisierungen und Texte kombinieren.

**Visualisierungswerkzeuge:**

o Matplotlib, Seaborn und andere Bibliotheken zur Erstellung statischer, animierter und interaktiver Visualisierungen, einschließlich Baumkarten, Akkorddiagrammen und 3D-Animationen.

**Projektbeschreibung:**

**Datenübernahme und Speicherung:**

o Das Projekt beginnt mit der Übernahme von Schulungs- und Test-CSV-Dateien von Kaggle in das Hadoop-Ökosystem unter Verwendung von HDFS auf einem lokalen Rechner.

o Zusätzliche CSV-Dateien, wie Abonnementdaten, werden ebenfalls in HDFS übernommen und gespeichert, um eine effiziente Abfrage und weitere Verarbeitung zu ermöglichen.

**Datenverarbeitung mit Hadoop MapReduce:**

o Unter Verwendung des MapReduce-Paradigmas auf dem lokalen Rechner über WSL mit Ubuntu verarbeitet das Projekt die Testdaten-CSV, um erste Bereinigungs-, Transformations- und Aggregationsschritte durchzuführen.

o Die resultierende CSV aus dem MapReduce-Prozess wird erneut in HDFS gespeichert, um weitere Analysen durchzuführen.

**Datenbereinigung und -integration mit Pandas:**

o Die erste Datenbereinigung und -transformation werden unter Verwendung von Pandas auf den Schulungsdaten durchgeführt. Dies umfasst die Behandlung fehlender Werte, Filterung und Erstellung einer verarbeiteten CSV.

o Verschiedene Verknüpfungsoperationen (inner join, self join, left outer join, right outer join, full outer join) mit Abonnement-CSV-Dateien werden durchgeführt, um verschiedene Datenquellen mithilfe von Spark SQL zu integrieren.

**In-Memory-Datenverarbeitung mit Spark:**

o Spark wird verwendet, um die verarbeiteten und integrierten Daten aus HDFS zu laden.

o Fortgeschrittene Datenmanipulationen und Aggregationen werden unter Verwendung von Spark SQL durchgeführt.

o Die Bibliothek MLlib wird genutzt, um Regressionsanalysen und eingeschränkte Mustererkennungen anzuwenden, um Trends und Muster in den Daten zu identifizieren.

**Visualisierung und Analyse mit Jupyter Lab:**

o Die Ergebnisse der Spark-Berechnungen werden in Jupyter Lab importiert.

o Visualisierungsbibliotheken wie Matplotlib und Seaborn werden verwendet, um detaillierte und interaktive Plots zu erstellen, einschließlich **Baumkarten, Akkorddiagrammen und 3D-Animationen,** um die Daten zu repräsentieren.

o Diese Visualisierungen werden sich darauf konzentrieren, wie die Nutzung verschiedener sozialer Medienplattformen durch Männer und Frauen mit verschiedenen mentalen Gesundheitszuständen wie Glück, Langeweile, Traurigkeit und Angst korreliert.

Zukunftsaussichten:

o Erforschen Sie die Integration fortschrittlicherer maschineller Lernalgorithmen und KI-Techniken, um die Vorhersagefähigkeiten und Erkenntnisse aus den Daten weiter zu verbessern.

Erwartete Ergebnisse:

• Effiziente Verarbeitung und Speicherung großer Datensätze unter Verwendung von HDFS auf einem lokalen Rechner.

• Verbesserte Datenverarbeitungsgeschwindigkeit und -leistung unter Verwendung von Hadoop MapReduce und Apache Spark.

• Detaillierte und interaktive Visualisierungen bieten klare Einblicke in die Daten.

• Einblicke, wie sich verschiedene soziale Medienplattformen unterschiedlich auf die psychische Gesundheit von Männern und Frauen auswirken.

• Grundlage für die zukünftige Integration fortschrittlicher maschineller Lern- und KI-Techniken.