Woche 9 – Datenaufbereitung und Analyse mit Python

Einleitung

Willkommen zu einer faszinierenden Woche, die ganz im Zeichen der Datenaufbereitung und Datenanalyse unter Einsatz von Python und Pandas steht. In einer Welt, in der Daten eine immer wichtigere Rolle spielen, ist die Fähigkeit, Daten zu interpretieren und daraus Schlüsse zu ziehen, von unschätzbarem Wert. Es geht darum, Rohdaten in wertvolle Erkenntnisse zu verwandeln und die daraus resultierenden Informationen effektiv zu nutzen.

Erfahren Sie, wie Sie die Datenqualität beurteilen, mit fehlenden Werten umgehen und tiefgreifende statistische Analysen durchführen können. Diese Kenntnisse sind in einer Vielzahl von Branchen wie Marketing, Finanzen, Gesundheitswesen und Technologie von großer Bedeutung.

Diese Woche ist daher ein entscheidender Schritt für alle, die sich für die Bereiche der Datenwissenschaft und Datenanalyse begeistern oder datenbasierte Entscheidungen in ihrem beruflichen Umfeld treffen möchten.

Im Laufe dieser Woche konzentrieren Sie sich auf das Selbststudium und die praktische Anwendung. Entdecken Sie, wie Sie mit Python und Pandas Daten effizient laden, bereinigen und für die Analyse vorbereiten können. Diese Fähigkeiten sind unerlässlich für jede Form der Datenanalyse.

Sobald Sie die Grundlagen gemeistert haben, werden Sie sich fortgeschrittenen Techniken der Datenanalyse zuwenden. Dazu gehört auch die Erstellung aussagekräftiger Datenvisualisierungen, die komplexe Zusammenhänge leicht verständlich darstellen.

Der Präsenztag bietet Ihnen die Gelegenheit, Ihr neu erworbenes Wissen anzuwenden, offene Fragen zu klären und direktes Feedback zu erhalten. Wir werden reale Datensätze analysieren und praktische Herausforderungen bewältigen, um Ihr Verständnis und Ihre analytischen Fähigkeiten zu vertiefen.

Bis zum Ende dieser Woche werden Sie nicht nur die Fähigkeiten besitzen, Daten mit Python und Pandas effizient zu verarbeiten und zu analysieren, sondern auch ein tiefgreifendes Verständnis für die Bedeutung und die Kraft der Datenanalyse entwickeln. Sie werden darauf vorbereitet sein, in Ihrem Fachbereich fundierte, datenbasierte Entscheidungen zu treffen und die Geschichten zu entdecken, die in den Daten verborgen liegen.

Freuen Sie sich darauf, Ihre Kenntnisse im Bereich der Datenanalyse zu erweitern und zu vertiefen. Diese Woche wird herausfordernd, aber äußerst lohnend sein. Lassen Sie uns gemeinsam die Welt der Daten mit Python und Pandas erforschen!

Gliederung

Hier ein Überblick über die Inhalte und Aktivitäten der aktuellen Woche:

- Selbststudium:
    - Grundlagen und Installation der Pandas-Bibliothek
    - Übersicht über Pandas-Datenstrukturen: DataFrame und Series
    - Laden und Speichern von Daten in verschiedenen Formaten (CSV, Excel, JSON)
    - Datenbereinigung: Umgang mit fehlenden Daten, Duplikaten und Datentypkonvertierungen
    - Erkundung von Datensätzen mit grundlegenden statistischen Funktionen
    - Auswahl und Filterung von Daten
    - Gruppierung und Aggregation von Daten
    - Einführung in grundlegende Datenvisualisierungstechniken mit Pandas und Matplotlib
    - Erstellen von Diagrammen und Plots zur Darstellung von Datenzusammenhängen
- Aufgaben:
    - Durchführung spezifischer Datenanalyseaufgaben mit realen Datensätzen
    - Anwendung verschiedener Datenmanipulations- und Visualisierungstechniken
    - Entwicklung eines kleinen Projekts, das die Datenaufbereitung und -analyse in Python und Pandas umfasst
- Präsenztag:
    - Wiederholung
    - Praktische Anwendung von Pandas in Datenanalyse-Szenarien
    - Einführung in fortgeschrittene Datenmanipulationstechniken
    - Ausblick: Datenvisualisierungen

Inhalte und thematische Abgrenzung

Die folgende Auflistung zeigt detailliert, welche Themen Sie in der Woche behandeln und bearbeiten. Sie sind eine Voraussetzung für die folgenden Wochen und sollten gut verstanden worden sein. Wenn es Verständnisprobleme gibt, machen Sie sich Notizen und fragen Sie am Präsenztag nach, so dass wir gemeinsam zu Lösungen kommen können. Und denken Sie bitte immer daran: es gibt keine „dummen“ Fragen!

1. Einführung in Pandas:

    - Grundlagen der Pandas-Bibliothek: Installation und Übersicht.
    - Pandas-Datenstrukturen: DataFrame und Series.

2. Datenimport mit Pandas:

    - Laden von Daten aus verschiedenen Quellen (CSV, Excel, JSON).
    - Erste Schritte mit dem DataFrame.

3. Datenbereinigung:

    - Behandlung fehlender Daten und Duplikate.
    - Datentypkonvertierung und -normalisierung.

4. Datenexploration:

    - Grundlegende statistische Analyse von Daten.
    - Nutzung von Pandas-Funktionen zur Datenuntersuchung.

5. Datenmanipulation:

    - Auswahl und Filterung von Daten.
    - Gruppierung und Aggregation von Daten.

6. Datenvisualisierung:

    - Einführung in die Datenvisualisierung mit Pandas und Matplotlib.
    - Erstellen von Diagrammen und Plots zur Datenanalyse.

7. Fortgeschrittene Datenmanipulation:

    - Kombinieren, Verbinden und Umstrukturieren von Daten.
    - Anwendung komplexer Datenmanipulationstechniken.

8. Zeitreihenanalyse:

    - Arbeit mit Zeitstempeln und zeitlichen Daten.
    - Analyse von Zeitreihendaten mit Pandas.

Lernpfad

Der Lernpfad ist ein Vorschlag, in welcher Reihenfolge Sie die Inhalte der Woche angehen können. Betrachten Sie ihn gerne als eine Todo-Liste, die Sie von oben nach unten abhaken. So können Sie sicher sein, dass Sie alle wichtigen Themen bearbeitet haben und sind gut vorbereitet für die folgenden Wochen.

1. Einführung in Pandas und Datenstrukturen:

    - Beginnen Sie mit einer Einführung in die Pandas-Bibliothek und deren Kernkonzepte.
    - Erforschen Sie Pandas-Datenstrukturen wie DataFrame und Series.

2. Datenimport und -aufbereitung:

    - Lernen Sie, wie man Daten aus verschiedenen Quellen wie CSV, Excel und JSON in Pandas lädt.
    - Üben Sie Techniken der Datenbereinigung, einschließlich der Behandlung fehlender Daten und der Datentypkonvertierung.

3. Erkundung und Manipulation von Daten:

    - Erarbeiten Sie Methoden zur Erkundung von Datensätzen, wie z.B. die Anwendung statistischer Funktionen.
    - Üben Sie die Auswahl, Filterung und Gruppierung von Daten sowie verschiedene Aggregationsmethoden.

4. Datenvisualisierung:

    - Einführung in die Grundlagen der Datenvisualisierung mit Pandas und Matplotlib.
    - Erstellen Sie verschiedene Arten von Diagrammen und Plots, um Datentrends und Muster zu erkennen.

5. Praktische Anwendung von Pandas:

    - Wenden Sie die gelernten Techniken an, um spezifische Datenanalyseaufgaben mit realen Datensätzen zu lösen.
    - Führen Sie verschiedene Datenmanipulations- und Visualisierungstechniken durch.

6. Projektarbeit:

    - Entwickeln Sie ein kleines Projekt, das die Datenaufbereitung und -analyse in Python und Pandas umfasst.
    - Wählen Sie einen Datensatz und wenden Sie die gelernten Methoden an, um interessante Erkenntnisse zu gewinnen.

7. Vorbereitung auf den Präsenztag:

    - Wiederholen Sie die in der Woche gelernten Konzepte und bereiten Sie Fragen und Diskussionspunkte für den Präsenztag vor.
    - Nutzen Sie die Gelegenheit, um tiefer in die Materie einzutauchen und Ihr Verständnis zu festigen.

Programmieraufgaben

Die folgenden Programmieraufgaben sollen Ihnen eine Anregung geben. Haben Sie eigene Ideen und Themen, die Sie ausprobieren wollen, dann sollten Sie diesen nachgehen. Wichtig ist vor allem, dass Sie „Dinge ausprobieren“. Und auch, dass Sie Fehler machen, sowohl syntaktische als auch semantische. Versuchen Sie diese Fehler zu finden und aufzulösen, dann gerade aus den Fehlern lernen Sie am Ende am meisten.

1. Explorative Datenanalyse mit Pandas:

Untersuchen Sie einen bereitgestellten Datensatz (z.B. Finanz-, Wetter- oder soziale Medien Daten) mit Pandas. Laden Sie den Datensatz in einen DataFrame, führen Sie eine explorative Datenanalyse durch, indem Sie statistische Zusammenfassungen, Sortierungen, Filterungen und Gruppierungen anwenden.

2. Datenbereinigung und -transformation:

Verbessern Sie die Qualität eines realen, „unordentlichen“ Datensatzes. Identifizieren und behandeln Sie fehlende Werte, entfernen Sie Duplikate, konvertieren Sie Datentypen und führen Sie Datentransformationen durch (z.B. Normalisierung, Binning).

3. Visualisierung von Datenbeziehungen:

Erstellen Sie aussagekräftige Visualisierungen, um Beziehungen in einem Datensatz zu erkennen. Verwenden Sie Pandas zusammen mit Matplotlib oder Seaborn, um verschiedene Arten von Diagrammen (wie Scatterplots, Histogramme, Boxplots) zu erstellen.

4. Datenaggregation und Gruppierung:

Erlernen Sie fortgeschrittene Gruppierungs- und Aggregationstechniken. Gruppieren Sie Daten nach verschiedenen Kriterien und wenden Sie komplexe Aggregationsfunktionen an.

5. Erstellen von Datenpipelines:

Entwickeln Sie eine automatisierte Datenpipeline. Automatisieren Sie den Prozess der Datenaufbereitung, -analyse und -berichterstattung.

Abschluss-Quiz

Das Quiz soll Ihnen einen ersten Hinweis auf Ihren Lernfortschritt geben. Nach unserer Einschätzung sollten Sie diese Fragen alle beantworten können, wenn Sie den Stoff der Woche durchgearbeitet und verstanden haben. Natürlich gibt es noch sehr viel mehr mögliche Fragen, dazu wollen wir auf die Literatur und das Internet verweisen. Geben Sie gerne einmal „python quizzes“ bei Google ein.

1. Was ist der Hauptvorteil der Verwendung von Pandas für die Datenanalyse in Python?

   a) Erhöhte Ausführungsgeschwindigkeit von Programmen.

   b) Automatische Visualisierung von Daten.

   c) Vereinfachte Handhabung und Analyse von Datensätzen.

   d) Integrierte KI-Fähigkeiten zur Datenanalyse.

2. Welche Methode wird in Pandas verwendet, um fehlende Daten in einem DataFrame zu ersetzen?

   a) replace()

   b) fillna()

   c) dropna()

   d) transform()

3. Was ermöglicht die Funktion groupby() in Pandas?

   a) Gruppierung von Daten basierend auf einer oder mehreren Spalten.

   b) Verschlüsselung der Daten in einem DataFrame.

   c) Sortierung des DataFrames nach Index.

   d) Konvertierung eines DataFrame in ein anderes Format.

4. Für welche Aufgabe wird die Bibliothek Matplotlib hauptsächlich verwendet?

   a) Datenbereinigung.

   b) Maschinelles Lernen.

   c) Datenvisualisierung.

   d) Webentwicklung.

5. Welche Aussage über merge() in Pandas ist korrekt?

   a) Es wird verwendet, um Spalten in einem DataFrame zu löschen.

   b) Es kombiniert Daten aus verschiedenen DataFrames basierend auf gemeinsamen Spalten.

   c) Es dient dazu, Datenreihen in einem DataFrame zu mischen.

   d) Es wird verwendet, um Daten aus einem DataFrame in eine Datenbank zu exportieren.

6. Welche Funktion lädt Daten aus einer CSV-Datei in einen Pandas DataFrame?

a) pandas.load_csv()

b) pandas.open_csv()

c) pandas.read_csv()

d) pandas.get_csv()

7. Wie kann man in Pandas die ersten fünf Zeilen eines DataFrames anzeigen?

a) df.first(5)

b) df.head(5)

c) df.begin(5)

d) df.start(5)

8. Was beschreibt die Funktion pivot_table() in Pandas?

a) Sie dreht die Ausrichtung eines DataFrames um 90 Grad.

b) Sie erstellt eine Zusammenfassungstabelle aus einem DataFrame.

c) Sie transformiert die Datenstruktur eines DataFrames in eine Tabelle.

d) Sie teilt einen DataFrame in mehrere kleinere DataFrames.

9. Wie kann man in Pandas eine neue Spalte zu einem DataFrame hinzufügen?

a) Mit der Methode append().

b) Durch direkte Zuweisung mit dem Spaltennamen.

c) Mit der Methode insert().

d) Durch Nutzung der Funktion concat().

10. Was ist ein wichtiger Schritt bei der Datenbereinigung mit Pandas?

a) Das Programmieren von Machine Learning-Modellen.

b) Die Erstellung von Pivot-Tabellen zur Datensynthese.

c) Die Identifizierung und Behandlung von fehlenden oder inkonsistenten Daten.

d) Das Schreiben von SQL-Abfragen innerhalb von Python.

Ressourcen

Als Ressourcen sind die Online-Dokumentationen der verwendeten Frameworks zu nutzen.