Project Summary

College Football Coach Salary Recommendations

github repo

This was a key deliverable in IST 718, Big Data Analytics. It demonstrates an advanced level of data manipulation and analysis, involving a series of complex tasks, including scraping and fuzzy matching to link records. The resulting dataset was used to model the salaries of college coaches.

Overview

Objective

To execute a comprehensive data analysis process, starting from data collection to modelling.

Challenge

Matching records from 6 different sources.

Data Source

Data was acquired through scraping, API calls, and from open-source websites like Wikipedia.

Technical Process

Data Scraping and Preprocessing

Utilized Python libraries including requests, lxml.html, and pandas for data scraping and preprocessing.

Fuzzy Matching Functions

Employed fuzzy matching techniques to identify and correct mismatches in data. This included the use of the libraries thefuzz and Levenshtein for efficient string matching.

Exploratory Data Analysis

Conducted an in-depth exploratory analysis to uncover insights from the data, ultimately focusing on choosing variables to include in model.

Modeling

Created 9 different linear models to recommend college coach salaries, and selected the best-performing model.

Challenges and Problem-Solving

Overcoming the challenges of integrating data from diverse sources, ensuring consistency and accuracy.

Tackling the intricacies of fuzzy matching to align disparate data sets accurately.

Insights Gained

The project highlighted the importance of meticulous data cleaning and preparation in data analysis.
Gained practical experience in using advanced Python libraries for data processing tasks.

Reflection on Data Science Techniques

Completing this provided in-depth experience in applying advanced data manipulation techniques, crucial for extracting meaningful insights from complex datasets.
It also underlined the significance of precision in data preprocessing, which is foundational for accurate data analysis and modeling.