Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Informatyki

Studia Podyplomowe
Big Data - przetwarzanie i analiza dużych zbiorów danych

PRACA KOŃCOWA

Paweł Safuryn

Wykorzystanie analizy big data na statystykach meczów tenisowych w celu poprawy przygotowania i wyników zawodnika

Opiekun pracy
mgr inż. Damian Warszawski

Warszawa, 2024

STRESZCZENIE

Analityka danych w tenisie ziemnym jest tematem stosunkowo nowym i nierozwiniętym pomimo ogromnej ilości szczegółowych danych zbieranych podczas meczów profesjonalnych rozgrywek. Celem pracy jest zbadanie potencjału analizy ogólnodostępnych statystyk meczowych w celu poprawy przygotowania i wyników zawodnika. Surowe dane (262 pliki CSV o całkowitym rozmiarze 395,7 MiB) zawierające wysokopoziomowe statystyki meczów zostały pozyskane z Internetu. Dane zostały oczyszczone, przetworzone i przeanalizowane z użyciem narzędzi big data takich jakich AWS CLI, Amazon S3, AWS Glue, Amazon EMR, czy PySpark. Cały kod został udostępniony w repozytorium na GitHubie (https://github.com/safurynp/pw-big-data-final-project). Praca analityczna skupiła się na zrozumieniu wpływu rodzaju tenisowej nawierzchni na wyniki zawodnika, analizie serwisu na różnych poziomach rankingu i wśród zawodników o różnym wzroście oraz analizie wpływu zmęczenia na wyniki zawodników. Użyteczność analizy tenisowych danych została wykazana pomimo użycia stosunkowo małego zbioru danych w kontekście big data. Rozszerzenie tego zbioru i przeprowadzenie analiz na większych i bardziej różnorodnych danych (np. historia meczu punkt po punkcie, nagrania wideo meczów, informacje z systemu Hawk-Eye) ma potencjał zrewolucjonizować podejście trenerów, organizacji i samych zawodników do ich tenisowego rozwoju.

Słowa kluczowe: tenis, big, data, analiza

Using big data analysis on statistics of tennis matches to improve player preparation and performance

Data analysis in tennis is a relatively new and underdeveloped topic despite huge amount of detailed data collected during professional tennis matches. The goal of this work is exploring the potential of analysing widely-available tennis match statistics in order to improve player preparation and performance. Raw data (262 CSV files with total size of 395,7 MiB) containing high-level match statistics were obtained from the internet. The data was cleaned, transformed and analysed with the use of big data tools like AWS CLI, Amazon S3, AWS Glue, Amazon EMR or PySpark. All code was made publicly available on a GitHub repository (https://github.com/safurynp/pw-big-data-final-project). The analytical work focused on understanding the effect of tennis surface on player performance, analysing serve quality for players with different ranking and height, and analysing how player tiredness affects their performance. The usefulness of tennis data analysis has been demonstrated despite the use of a relatively small dataset in the context of big data. Expanding this dataset and conducting analyses on larger and more diverse datasets (such as point-by-point match history, video recordings of tennis matches, data from the Hawk-Eye system) has the potential to revolutionise the ways in with coaches, organisations, and players themselves approach their tennis development.

Keywords: tennis, big, data, analysis