SCS 3252-010 – PROJECT

Steven chiu

clustering nba players with apache spark & mllib

by Steven Chiu

2018

Contents

[Summary 2](#_Toc531479804)

[Project overview 2](#_Toc531479805)

[Data Sources 2](#_Toc531479806)

[Clustering Development using Databricks (Apache Spark + MLLIB & Pipelines) 3](#_Toc531479807)

[Design 3](#_Toc531479808)

[Development 3](#_Toc531479809)

# Summary

Professional sports is a multi-billion dollar industry and the resources (i.e players) need to be optimally used in order to maximize profit for the companies – sports clubs.

In this exercise, we will attempt to cluster various types of players using available data and determine the players that are competing at par with other players in the league.

# Project overview

This project will use the tools that is part of SCS-3252’s course curriculum in order to achieve the goal of clustering various players in a given set of data.

The technologies used are as follows:

|  |  |
| --- | --- |
| Databricks | Pyspark will be used for the project within databricks. Databricks already have pre-built clusters to interface with Apache Spark and pyspark environment complete with MLLIB. |
| Pyspark | Python + Spark will be used as code. |
| CSV data format | The data sources that have been researched for, will output their data in CSV format. Various sources have been looked at – nba.com, thescore.com, basket-ball reference.com.  Basketball-reference.com has been selected to be used for the statistics of players for the year to be downloaded. |
| GitHub URL | https://github.com/schiu73/SCS3252\_Project\_NBA |

# Data Sources

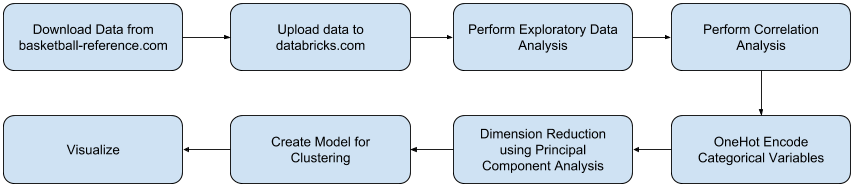
Basketball-reference.com: comprehensive site on NBA basketball statistics. The players accumulated statistics is easily accessible and downloadable via web browser.

|  |  |
| --- | --- |
| Game stats URL | https://www.basketball-reference.com/leagues/NBA\_2019\_per\_game.html |
|  |  |

# Clustering Development using Databricks (Apache Spark + MLLIB & Pipelines)

## Design

The following is the workflow on from source data to clustering via machine learning libraries to cluster similar NBA players together.



## Development

|  |  |
| --- | --- |
| Technology | Purpose |
| Databricks.com | Clusters preloaded with Apache Spark and PySpark. Environment used for development |
| URL |  |
|  |  |

Notebook URL: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1793597675784580/3496336848243693/8868352524310076/latest.html>

Presentation Slide deck:

<https://github.com/schiu73/SCS3252_Project_NBA/raw/master/SCS_3252_010_Steven_Chiu_Project.pptx>