

Web Scraping: Top Rated Indian Movies

Introduction

This is a beginner project where I scraped data on top rated Indian movies from the IMDB movie website, <http://www.imdb.com/> (<http://www.imdb.com/>) and put it in a pandas dataframe for analysis. Simple analysis is done to practice and revise simple concepts.

Method

I used Python 3.6.3 and jupyter notebook for the project. Beautiful Soup and requests libraries were used to scrape the data from IMDB movie website. Request library was used to request the web page and Beautiful soup to go through the html code and pull out the attributes. I stored the data in Python lists and merged the list together to make a pandas dataframe. Next step I checked for any discrepancies. Year attribute was within parentheses which was removed. Then I did a simple analysis using libraries pandas, numpy and seaborn.