acardocacho / DSI_LDN_1_LESSON_NOTES Private
forked from ga-students/DSI_LDN_1_LESSON_NOTES

Unwatch 5 Star 0 Fork 26

Code Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master DSI_LDN_1_LESSON_NOTES / projects / project-06 /

Create new file Upload files Find file History

This branch is 63 commits ahead, 2 commits behind ga-students:master. Pull request Compare

aidanrussell project 6 Latest commit 1e0b222 a day ago

..

project-06-rubric.md	project 6	a day ago
readme.md	project 6	a day ago

readme.md

Project 6: APIs + Random Forests

Overview

This week, we learned about ensemble methods, APIs, and natural language processing. Now we're going to put these skills to the test. You've been hired by Netflix to examine what factors lead to certain ratings for movies. Given that Netflix does not currently store this type of data, your boss has suggested that you collect ratings and reviews data from IMDB. Netflix is no stranger to machine learning, however:

- Netflix uses random forests and decision trees to predict what types of movies an individual user may like.
- Using unsupervised learning techniques, they are able to continually update suggestions, listings, and other features of it's user interface.
- Netflix, however, hasn't focused on collecting data on the top movies of all time, and would like to add some of them to their offerings based on popularity and other factors.

Point: Your boss isn't sure where to start on this project, so your task is to collect the data and construct a random forest to understand what factors contribute to ratings.

Project Summary

Acquire data from IMDB, and use whatever metrics you can collect to predict whether it is a good movie.

When you've finished your analysis, Netflix would like a report detailing your findings, with recommendations as to next steps.

Here are some questions to keep in mind:

- What factors are the most direct predictors of rating?
- You can use rating as your target variable. But it's up to you whether to treat it as continuous, binary, or multiclass.

Goal: Completed Jupyter notebook that includes modeling using a random forest and an blog post explaining your findings.

Requirements

This is deliberately open ended. There is no starter code. It's up to you how to acquire the data, store the data, and what

features you want to use.

We expect you to use a **tree-based model**, but the rest of the decisions are up to you.

We will be looking for the following things:

- A clear problem statement & description of the goals of your study to be included in the final report
- Data from IMDB
- Cleaned and refined data
- Visualization. Plots that describe your data and evaluate your model.
- Tree-based models (use any combination of ensemble techniques: random forests, bagging, boosting).
- A blog post presenting the results of your findings as a report to Netflix, including:
  - a problem statement,
  - summary statistics of the various factors (e.g. year, number of ratings, etc.),
  - your model,
  - at least 2 graphics,
  - and your recommendations for next steps!

Necessary Deliverables / Submission

- Materials must be in a clearly labeled Jupyter notebook
- Link to the blog post with your report in your Jupyter notebook
- Materials must be submitted to GitHub by Monday of Week 7

Suggested Ways to Get Started

- You can get data on the top 250 movies on IMDB using the IMBDpie API
- If you need additional data, you can either research additional APIs, or scrape it yourself using BeautifulSoup
- Read the docs for whatever technologies you use. Most of the time, there is a tutorial that you can follow, but not always, and learning to read documentation is crucial to your success!
- Document **everything**.

Useful Resources

Documentation for BeautifulSoup

Project Feedback + Evaluation

Attached here is a complete rubric for this project.

Your instructors will score each of your technical requirements using the scale below:

Score	Expectations
0	*Incomplete.*
1	*Does not meet expectations.*
2	*Meets expectations, good job!*
3	*Exceeds expectations, you wonderful creature, you!*

This will serve as a helpful overall gauge of whether you met the project goals, but **the more important scores are the individual ones** above, which can help you identify where to focus your efforts for the next project!