

AUFGABE1:

Über Datensatz

Kontext

Diese Dateien enthalten Metadaten für alle 45.000 Filme, die im vollständigen MovieLens-Datensatz aufgeführt sind. Der Datensatz besteht aus Filmen, die am oder vor Juli 2017 veröffentlicht wurden. Zu den Datenpunkten gehören Besetzung, Crew, Handlungsschlüsselwörter, Budget, Einnahmen, Poster, Veröffentlichungsdaten, Sprachen und Produktionsaktion Unternehmen, Länder, TMDB-Stimmenzahlen und Stimmendurchschnitte.

Dieser Datensatz enthält außerdem Dateien mit 26 Millionen Bewertungen von 270.000 Benutzern für alle 45.000 Filme. Die Bewertungen liegen auf einer Skala von 1 bis 5 und wurden von der offiziellen GroupLens-Website abgerufen.

Inhalt

Dieser Datensatz besteht aus den folgenden Dateien:

`movies_metadata.csv`: Die Hauptmetadatendatei für Filme. Enthält Informationen zu 45.000 Filmen, die im vollständigen MovieLens-Datensatz enthalten sind. Zu den Funktionen gehören Poster, Hintergründe, Budget, Einnahmen, Veröffentlichungsdaten, Sprachen, Produktionsländer und Unternehmen.

`keywords.csv`: Enthält die Filmplot-Schlüsselwörter für unsere MovieLens-Filme. Verfügbar in Form eines stringifizierten JSON-Objekts.

`credits.csv`: Besteht aus Besetzungs- und Crewinformationen für alle unsere Filme. Verfügbar in Form eines stringifizierten JSON-Objekts.

`links.csv`: Die Datei, die die TMDB- und IMDB-IDs aller im Full MovieLens-Datensatz enthaltenen Filme enthält.

`links_small.csv`: Enthält die TMDB- und IMDB-IDs einer kleinen Teilmenge von 9.000 Filmen des vollständigen Datensatzes.

`Ratings_small.csv`: Die Teilmenge von 100.000 Bewertungen von 700 Benutzern zu 9.000 Filmen.

Auf den vollständigen MovieLens-Datensatz, der aus 26 Millionen Bewertungen und 750.000 Tag-Anwendungen von 270.000 Benutzern für alle 45.000 Filme in diesem Datensatz besteht, kann hier zugegriffen werden

Danksagungen

Dieser Datensatz ist ein Ensemble von Daten, die von TMDB und GroupLens gesammelt wurden. Die Filmdetails, Credits und Schlüsselwörter wurden von der TMDB Open API gesammelt. Dieses Produkt nutzt die TMDb API, ist aber nicht von TMDb unterstützt oder zertifiziert. Ihre API bietet auch Zugriff auf Daten zu vielen weiteren Filmen, Schauspielern, Crewmitgliedern, und Fernsehsendungen. Hier können Sie es selbst ausprobieren.

Die Filmlinks und Bewertungen wurden von der offiziellen GroupLens-Website abgerufen. Die Dateien sind Teil des hier verfügbaren Datensatzes

Inspiration

Dieser Datensatz wurde im Rahmen meines zweiten Capstone-Projekts für den Data Science Career Track von Springboard zusammengestellt. Ich wollte eine umfassende EDA zu Filmdaten durchführen, um die Geschichte und die Geschichte des Kinos zu erzählen, und diese Metadaten in Kombination mit MovieLens-Bewertungen verwenden, um verschiedene Typen zu erstellen von Empfehlungssystemen.

Meine beiden Notizbücher sind als Kernel mit diesem Datensatz verfügbar: The Story of Film and Movie Recommender Systems

Einige der Dinge, die Sie mit diesem Datensatz tun können:








Vorhersage von Filmeinnahmen und/oder Filmerfolg auf der Grundlage einer bestimmten Kennzahl.

Welche Filme erhalten in der TMDb tendenziell höhere Stimmenzahlen und Stimmendurchschnitte?

Aufbau inhaltsbasierter und auf kollaborativer Filterung basierender Empfehlungsmaschinen.

Gegeben sind folgende CSV Dateien

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download>

 credits.csv	22.06.2023 21:02	Microsoft Excel-C...	185.467 KB
 keywords.csv	22.06.2023 21:02	Microsoft Excel-C...	6.086 KB
 links.csv	22.06.2023 21:02	Microsoft Excel-C...	966 KB
 links_small.csv	22.06.2023 21:02	Microsoft Excel-C...	180 KB
 movies_metadata.csv	22.06.2023 21:02	Microsoft Excel-C...	33.638 KB
 ratings.csv	22.06.2023 21:02	Microsoft Excel-C...	692.921 KB
 ratings_small.csv	22.06.2023 21:02	Microsoft Excel-C...	2.382 KB

Schreibe mittels Mrjob einen Mapreduce Algorithmus der die Top 10 best bewerteten Filme und führe es auf Hadoop aus. Speicher das Ergebnis in einer Datei als output.txt. Aus der Ergebnisliste in der output.txt Datei, schreibe ein Python Script der die Namen von den Top 10 best bewerteten Filme und deren Beschreibungen auflistet.

AUFGABE 2:

Lade die Amazon movie Reviews aus der Stanford universität (8M reviews von 889K Usern 3.1 GB) und überlege dir wie du die Top-10 meist bewerteten Filme Mittels eines Mapreduce Algorithmus auf Hadoop HDFS herausfindest.

<https://snap.stanford.edu/data/web-Movies.html>



[SNAP for C++](#) ▶
[SNAP for Python](#) ▶
[SNAP Datasets](#) ▶
[BIOSNAP Datasets](#)
[What's new](#)
[People](#)
[Papers](#)
[Projects](#) ▶
[Citing SNAP](#)
[Links](#)
[About](#)
[Contact us](#)

Open positions

Open research positions in **SNAP** group are available at [undergraduate](#), [graduate](#) and [postdoctoral](#) levels.

Web data: Amazon movie reviews

Dataset information

This dataset consists of movie reviews from [amazon](#). The data span a period of more than 10 years, including all ~8 million reviews up to October 2012. Reviews include product and user information, ratings, and a plaintext review. We also have reviews from [all other Amazon categories](#).

Dataset statistics

Number of reviews	7,911,684
Number of users	889,176
Number of products	253,059
Users with > 50 reviews	16,341
Median no. of words per review	101
Timespan	Aug 1997 - Oct 2012

Source (citation)

- J. McAuley and J. Leskovec. [From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews](#). WWW, 2013.

Files

File	Description
movies.txt.gz	Amazon movie data (~8 million reviews)

Data format

```

product/productId: B00006HAXW
review/userId: A1RSDE90N6RSZF
review/profileName: Joseph M. Kotow
review/helpfulness: 9/9
review/score: 5.0
review/time: 1042502400
review/summary: Pittsburgh - Home of the OLDIES
review/text: I have all of the doo wop DVD's and this one is as good or better than the
1st ones. Remember once these performers are gone, we'll never get to see them again.
Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE
this DVD !!
  
```