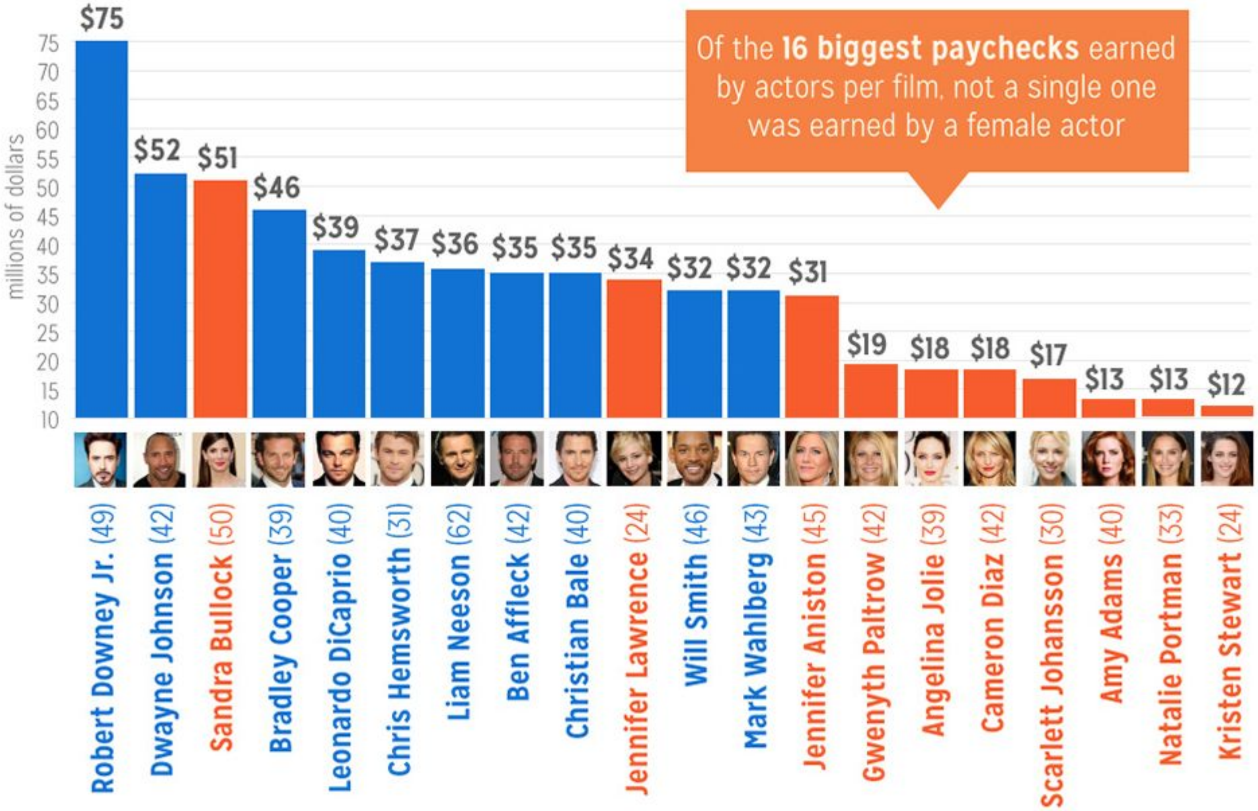
Technology Review:
Gender Diversity in Film & Negative Online Backlash

Trent Dillon, Rachel Franz, Eric Gomez, Jamie Park, Kate Van Ness



Of the **16 biggest paychecks** earned by actors per film, not a single one was earned by a female actor

source: New York Film Academy, 2014

+

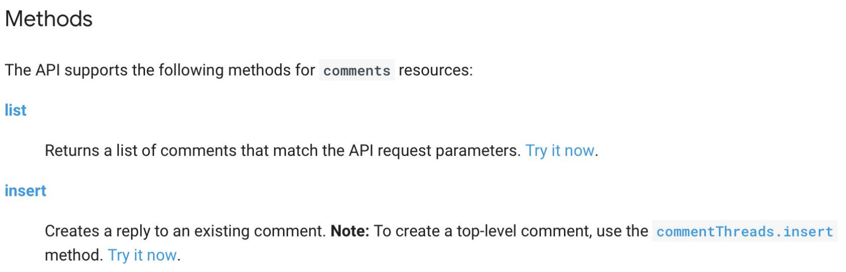
IMDb

↓

Are online commenters biased against movies that feature women?

YouTube Comment Scraper

Problem: How do we collect the entire list of online comments for 5000 movie trailers on YouTube?



Methods

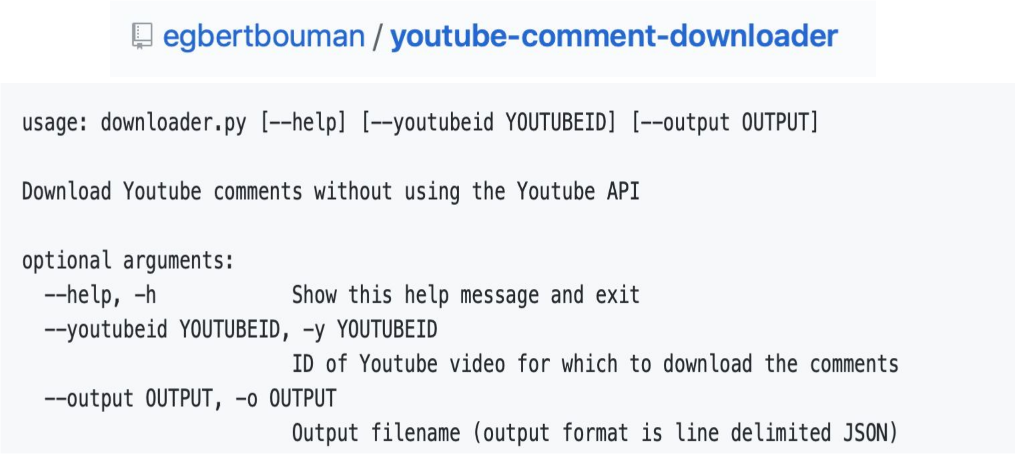
The API supports the following methods for `comments` resources:

list

Returns a list of comments that match the API request parameters. Try it now.

insert

Creates a reply to an existing comment. **Note:** To create a top-level comment, use the `commentThreads.insert` method. Try it now.



egbertbouman / **youtube-comment-downloader**

```
usage: downloader.py [--help] [--youtubeid YOUTUBEID] [--output OUTPUT]

Download Youtube comments without using the Youtube API

optional arguments:
  --help, -h            Show this help message and exit
  --youtubeid YOUTUBEID, -y YOUTUBEID
                        ID of Youtube video for which to download the comments
  --output OUTPUT, -o OUTPUT
                        Output filename (output format is line delimited JSON)
```

Pros: Well-documented
Cons: Complicated (e.g., API key), limited query search, have to write our own program, doesn’t automatically retrieve all comments

Pros: Easy to use, avoids API, returns all comments
Cons: Have to input video IDs (requires an additional program to retrieve our 5k movie trailers’ IDs)

Sentiment Analysis

TextBlob: built-in text processing python library

- How it works:
  - Input string, uses previously downloaded training data to analyze text
  - Outputs polarity score -1 to +1 and subjectivity score 0 to 1
- Appeal: easy to use, provides training data for download, language detection and translation
- Drawbacks: ability to process text abbreviations/emojis/etc common to short YouTube comments, memory issues with using your own training data

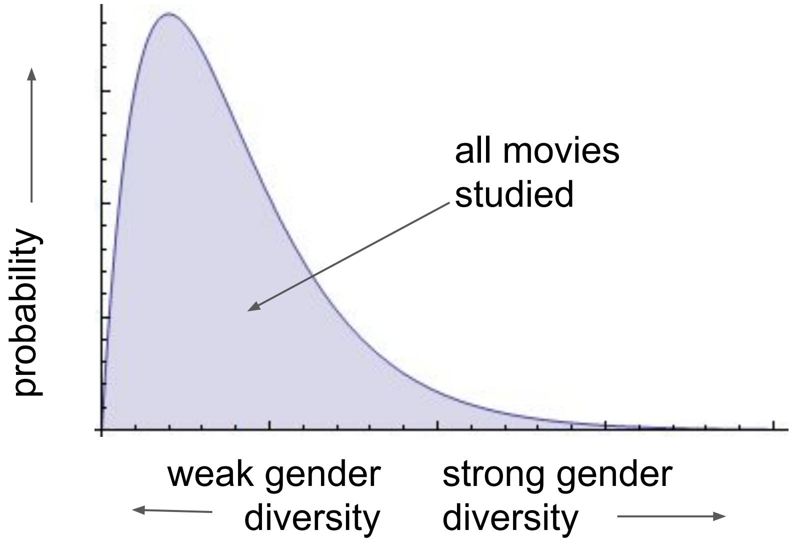
Sentiment Analysis

1) Nltk (Natural Language Toolkit) Vader
    a) Pre-trained classifier on messy social media posts
    b) Low accuracy: .53
2) Nltk Naive Bayes
    a) Trained on 160 comments
    b) Tested on 40 comments
    c) High accuracy: .83
    d) But probably overfitting the training data
    e) Have to label a large and representative training set

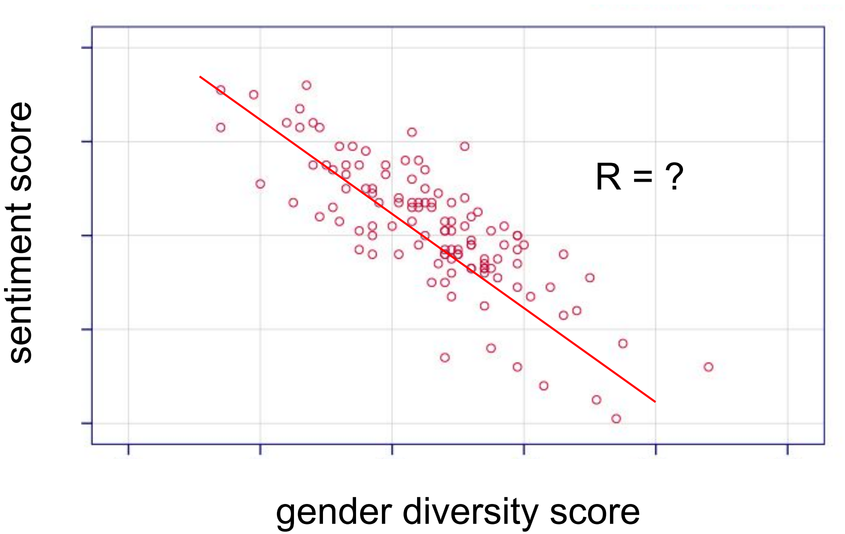
Visualization

results are likely to include two key figures...

Gender Diversity Score Probability Distribution



Correlation Visual



Visualization

we will produce unknown supplemental figures, but we really only expect to need one visualization library...

comprehensive statistical analysis…

- scatter plots
- bar/histogram plots
- distribution pots

also…

- very customizable
- interfaces well with the python data science stack (e.g. pandas and numpy libraries)