A Shallow Dive into the Deep Pool of Amazon Review Data

Zach Wright

4/30/2018

Visualize the Dive

NY Times API and UCSD Amazon Data

The Theory: A book listed on the NY Times Best Sellers list will have positive review sentiment

- The New York Times Best Seller list is widely considered the preeminent list of best-selling books in the United States.
- Amazon.com is recognized as the largest hub of book sales and offers thousands of user reviews per publication

The Plan:

-Conduct sentiment analysis on the Amazon reviews of the books most-frequently featured on the NYT Best Seller's list

Don't dive if you can't see the bottom

Flaws in design

Assumptions:

- The UCSD Amazon data is a collection of workable review data from 1996 - 2014
- I would access the NY Times Best Sellers History API and collect the data on these years

Reality:

- The UCSD Amazon data is contained in a 4GB JSON file, containing ∼9 million reviews
- NY Times API only provides recent Best Seller Data

Close your eyes and jump

Make it work

Attempts:

- Scrape the NYT Best Seller wikipedia pages for 1996 - 2014
- Open the 4GB JSON file in terminal, use grep() to find reviews based on the book's unique Amazon ID

Result:

- The JSON file only contains a few, if any, reviews of the books I'm interested in
- The JSON data is very dirty; missing characters, titles, Amazon ID's

Realization: This is not a viable approach for this analysis

Just Keep Swimming

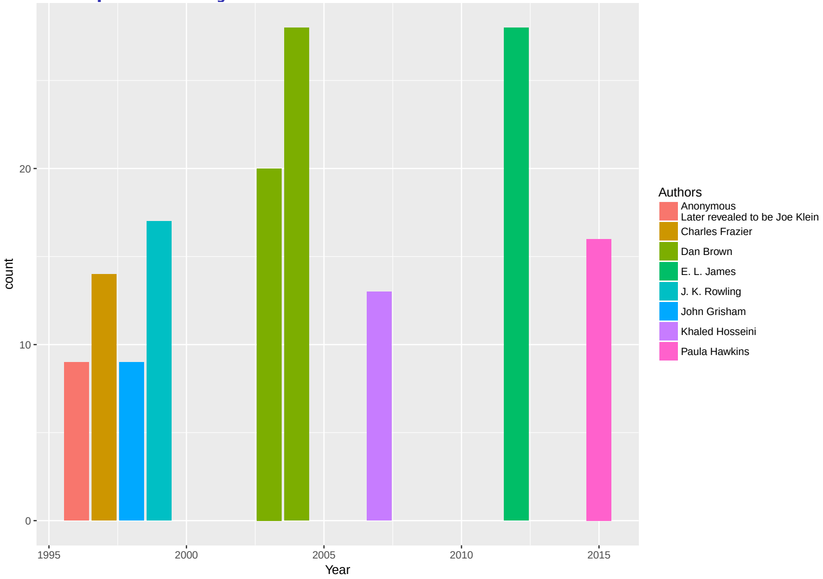
Try Smarter, Not Harder

Scrape the NYT Best Seller data:

- Who are the most frequent Authors?
- What are the most frequent Titles?

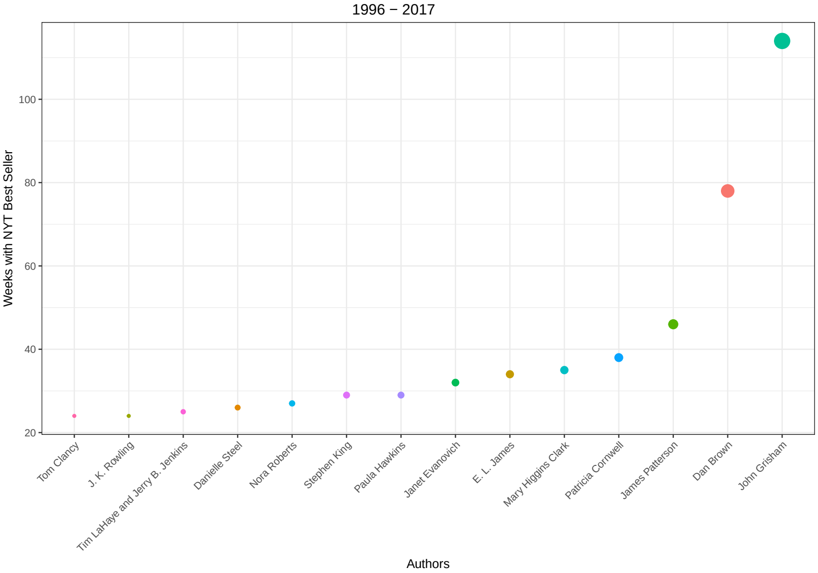
From there we can find which book reviews to analyze

NYT Graph Analysis



```
## $y
## [1] "Weeks"
```

NYT Graph Analysis

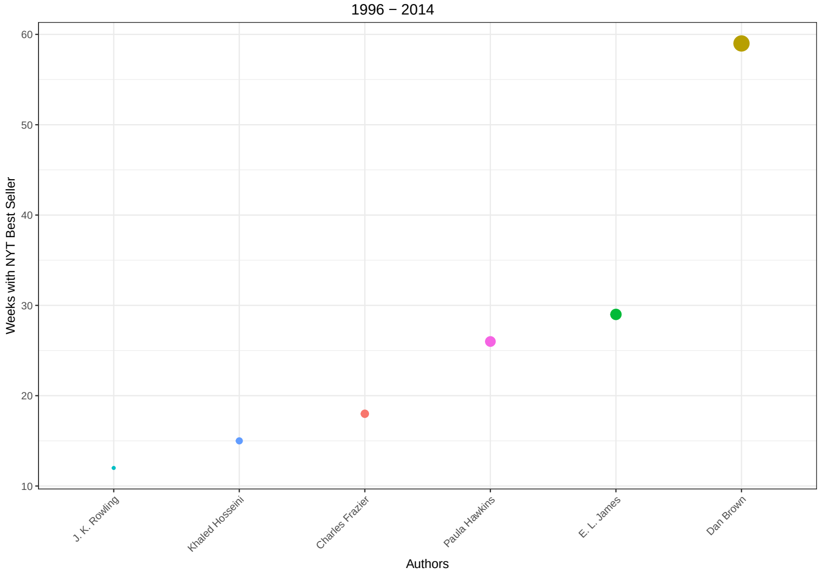


NYT Graph Analysis

Table 1: Most Frequent Best Seller Books 1996 - 2017

Book	Author	Weeks
The Da Vinci Code	Dan Brown	59
Fifty Shades of Grey	E. L. James	29
The Girl on the Train	Paula Hawkins	26
Cold Mountain	Charles Frazier	18
A Thousand Splendid Suns	Khaled Hosseini	15
Harry Potter and the Chamber of Secrets	J. K. Rowling	12

NYT Graph Analysis



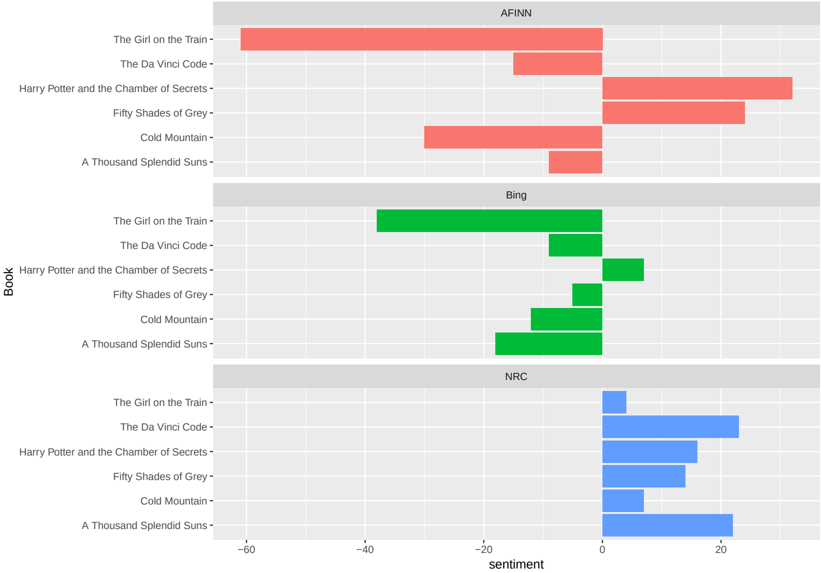
Access Usable Amazon Review Data

Scrape the initial Amazon Reviews for the most frequent Titles:

- Tidy the data
- Conduct sentiment analysis

See if the sentiment complements the Best Seller Status

Amazon Analysis



Conclusions

What do we think?

- This theory has not been proven
- Positive sentiment does not seem to be related to Best Seller reviews

What else could we do?

- Compare review sentiment to average Amazon Start score rating
- Scrape more reviews, analyze the sentiment over time
- Understand book reviews are subjective and difficult to analyze
- “This book was terrifying” could be considered a positive review for a thriller