An Analysis of Amazon Reviews

Joao Carreira
Outline

- Dataset and Methodology
- Sanity checks
- Dataset Analysis
  1. Characterization
  2. Products
  3. Users/Reviews
Dataset - Overview

- Amazon founded in 1994
- Amazon reviews 1995-2013 (18 year span)
- 34M reviews, 7M users, 2M products
- 35Gb of uncompressed data
- Dataset is available for research purposes [1]
- An analysis of review text is available [2]

[1] https://snap.stanford.edu/data/web-Amazon.html
C Programming in the Berkeley Unix Environment

This has been one of those books that I constantly refer to. Not only is it good for learning some of the unique C things that apply to Unix, but you can also learn how to get around in Unix. This is the book I learned C from, and it's still one of the first ones I go to when I need to refresh my brain about something.
1. Product Brand

B0000C2LFS Gifted Horse

2. Product Categories

0131097601
Books, Computers & Technology, Microsoft, Development, C & C++ Windows Programming
Books, Computers & Technology, Programming, APIs & Operating Environments, Unix
Books, Computers & Technology, Programming, Languages & Tools
Books, Computers & Technology, Software
Books, Education & Reference
Books, Science & Math, Mathematics

3. Product description

product/productId: 1878972405
product/description: Portuguese author Fernando Pessoa (1888-1935) published little in his lifetime, but his rediscovery in the 1990s has been as central to postmodernism as the rediscovery of Kafka in the 1950s was to modernism.

4. Related products

B000K85RMI also purchased 0684803305 0805062904
Methodology

- Exploratory analysis of the dataset
- This analysis focuses on products and users
- No textual analysis - NLP - of reviews
- Perl + R
- Code, graphs, and slides available @
  
github.com/jcarreirajamaz-study
# Sanity Checks

<table>
<thead>
<tr>
<th>Sanity Check</th>
<th>Description</th>
<th>Check?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct timestamps</td>
<td>Time between 95 and ‘13</td>
<td>✗</td>
</tr>
<tr>
<td>Helpfulness &lt;= 1</td>
<td>Helpfulness factor at most 1</td>
<td>✗</td>
</tr>
<tr>
<td>Price</td>
<td>Price is positive (and reasonable)</td>
<td>✗</td>
</tr>
<tr>
<td>Score 1-5</td>
<td>Score is a 1-5 value</td>
<td>✓</td>
</tr>
<tr>
<td>Review entries complete</td>
<td>All reviews have all entries</td>
<td>✗</td>
</tr>
<tr>
<td>Product price fluctuation</td>
<td>Different reviews for the same product may have different prices</td>
<td>✗</td>
</tr>
<tr>
<td>Review product title consistency</td>
<td>Review product title matches product title</td>
<td>✓</td>
</tr>
<tr>
<td>Daily activity cycle</td>
<td>Less reviews during night and more during day</td>
<td>✗</td>
</tr>
<tr>
<td>Products categories</td>
<td>All products have categories</td>
<td>✗</td>
</tr>
</tbody>
</table>
Sanity Checks

- Timestamps: Some are missing (e.g., “-1” entries)
- Timestamp hour at 4pm or 5pm
- Helpfulness: Some factors are > 1
  
  - product/productId: 1930771142
  - product/title: You Can Have Your Cheese and Eat It Too!
  - product/price: unknown
  - review/userId: A1VYC3XNQU72RF
  - review/profileName: William Cottringer
  - review/helpfulness: 2/1

- Price: Some products have price 0$. Others “unknown”
- Product price: prices are constant through time — not what happens in reality
- Some reviews do not have text (just summary)
- Some products have no category
Dataset Characterization

• How many reviews are made per year?
• What are the “biggest” products in amazon?
• How much do products cost?
• What are the most expensive categories?
• How often do users review products?
Reviews per Year

 Histogram of Reviews per Year

Frequency

Year

0e+00  1e+06  2e+06  3e+06  4e+06

1995  2000  2005  2010
Product Categories

Number of Products per Category

Product Count

Movies & TV
Music
Books
Clothing & Accessories
Health & Personal Care
Sports & Outdoors
Toys & Games
Home Improvement
Amazon Instant Video
Beauty
Automotive
All Electronics
Baby
Errors
Musical Instruments
Grocery & Gourmet Food
Industrial & Scientific
Cell Phones & Accessories
Software
Watches
Home Improvement
Art
Baby
Appliances
Office Supplies
Kindle Store
Office & School Supplies
Magazines
Unknown
Cameras & Photo
Computers
Product Prices

- Most products cost < 50$
- Prices capped at 999.99$

CDF of Product Prices
Product Prices

Boxplot of Price of Products by Category

- Outliers ignored
- Purchase circles - bestsellers lists for specific groups
Users Reviews

> 80% of users do not review more than 5 times
## Products - Questions

<table>
<thead>
<tr>
<th>Subject</th>
<th>Question</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life Expectancy</td>
<td>What is the life expectancy of a product?</td>
<td>Strong variations</td>
</tr>
<tr>
<td></td>
<td>Do reviews affect the life expectancy of products?</td>
<td>Probably</td>
</tr>
<tr>
<td></td>
<td>Do product life expectancy varies per product category?</td>
<td>Yes (e.g., books vs technology)</td>
</tr>
<tr>
<td>Reviews</td>
<td>Do review scores decay over time?</td>
<td>Depends on product category</td>
</tr>
<tr>
<td></td>
<td>Do reviews cluster at specific times (e.g., product launch)?</td>
<td>Should follow curve of adoption</td>
</tr>
</tbody>
</table>
Products - Life Expectancy

- Life expectancy: average number of years of life
- Considered only products with
  - > 50 reviews (frequently reviewed products)
  - last review before 2010 (no review likely means the product ‘died’)
- This filters reviews down to 4K products
Products - Life Span

Histogram of Product Life Span

CDF of Product Life Span
Products - Scores vs Life Expectancy

Correlation coefficient = 0.22 -> Scores do not affect life expectancy
Product Life Expectancy by Category

Boxplot of Life Expectancy of Products by Category

- Cross-classification of books and kindle
Review Scores Decay

- Compute the average decay of review scores over the years
- For each product scores are normalized to the first year average score
- Normalized scores are averaged per year after a product’s first review
- Products with less than 5 years of reviews and 3 reviews per year are ignored
- -> 28976 products
Review Scores Decay

Average Score per Year After First Review

Year After First Review

Normalized Average Score
Reviews Curve

- Compute reviews clustering throughout a product’s life — should follow curve of adoption
- For each product # of reviews is normalized
- # of reviews is averaged per year after a product’s first review
- Only “dead” products with no “holes” and at least 3 reviews per year considered
- -> 136 products
Reviews Curve

Normalized Number of Reviews per Year After First Review

Normalized Number of Reviews

Year After First Review
## User Reviews - Questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do users tend to review a product when they are either very satisfied or unsatisfied?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do positive / negative reviews tend to cluster in individual users, i.e., are there 'negative' users and 'positive' users?</td>
<td>Probably yes</td>
</tr>
<tr>
<td>Do users review products in a specific area of expertise or across different product categories?</td>
<td>Don’t know</td>
</tr>
<tr>
<td>Do users tend to be active reviewers over long periods of time?</td>
<td>No</td>
</tr>
<tr>
<td>What features of a review make it helpful?</td>
<td>Probably user experience and reviewer depth</td>
</tr>
</tbody>
</table>
Users - Scores

- Most reviews are positive
Users - Positive vs Negative Reviews

- Users with less than 10 reviews not considered
- Many “positive” users
Are Reviewers (1 Cat.) Experts?

- Check how many reviews are focused on a single category for each reviewer
- Ignore reviewers with less than 5 reviews
Are Reviewers (1 Cat.) Experts?
Users Life Expectancy

Histogram of Users Life Span

Time (Days)

Frequency

0 50000 150000 250000
Reviews Size vs Helpfulness

Correlation coefficient = 0.24
Reviewer Experience vs Helpfulness

- Correlation coefficient = -0.041
Questions?

Dataset - Review Records

- productID: 0131097601
- productTitle: C Programming in the Berkeley Unix Environment
- productPrice: $24.95
- reviewID: A1YJX5Q8Q6G02W
- reviewProfileName: Eugene Mab "physics Geek"
- reviewHelpfulness: 0/0
- reviewScore: 4.0
- reviewTime: 06/29/1999
- reviewSummary: Incredibly thin on my computer bookshelves
- reviewText: This has been one of those books that I constantly refer to. Not only is it good for learning some of the unique C things that apply to Unix, but you can also learn how to get around in Unix. This is the book I learned C from, and it's still one of the first ones I go to when I need to refresh my brain about something.

Sanity Checks

<table>
<thead>
<tr>
<th>Sanity Check</th>
<th>Description</th>
<th>Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct timestamps</td>
<td>Time between 96 and '13</td>
<td>✗</td>
</tr>
<tr>
<td>Helpfulness &lt;= 1</td>
<td>Helpfulness factor at most 1</td>
<td>✗</td>
</tr>
<tr>
<td>Price</td>
<td>Price is positive (and reasonable)</td>
<td>✗</td>
</tr>
<tr>
<td>Score 1-5</td>
<td>Score is a 1-5 value</td>
<td>✓</td>
</tr>
<tr>
<td>Review entries complete</td>
<td>All review have all entries</td>
<td>✓</td>
</tr>
<tr>
<td>Review price consistency</td>
<td>Review price matches product price</td>
<td>✓</td>
</tr>
<tr>
<td>Review product title consistency</td>
<td>Review product title matches product title</td>
<td>✓</td>
</tr>
</tbody>
</table>

Products - Questions

<table>
<thead>
<tr>
<th>Subject</th>
<th>Question</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifetime</td>
<td>What is the lifetime of a product?</td>
<td>Few years</td>
</tr>
<tr>
<td></td>
<td>Do reviews affect the lifetime of products?</td>
<td>Probably</td>
</tr>
<tr>
<td></td>
<td>Do product lifetime varies per product category?</td>
<td>Yes (e.g., books vs technology)</td>
</tr>
<tr>
<td>Reviews</td>
<td>Do reviews decay over time?</td>
<td>Yes, users become more demanding</td>
</tr>
<tr>
<td></td>
<td>Do reviews cluster at specific times (e.g., product launch)?</td>
<td>Should follow curve of adoption</td>
</tr>
</tbody>
</table>

User Reviews - Questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do users tend to review a product when they are either very satisfied or unsatisfied?</td>
<td>Yes</td>
</tr>
<tr>
<td>Do positive / negative reviews tend to cluster in individual users, i.e., are there 'negative' users and 'positive' users?</td>
<td>Probably yes</td>
</tr>
<tr>
<td>Do users review products in a specific area of expertise or across different product categories?</td>
<td>Yes (Books vs Technology)</td>
</tr>
<tr>
<td>Do users tend to be active reviewers over long periods of time?</td>
<td>No</td>
</tr>
<tr>
<td>What features of a review make it helpful?</td>
<td>Probably user experience and review depth</td>
</tr>
</tbody>
</table>

• [github.com/jcarreira/amazon-study](https://github.com/jcarreira/amazon-study)