Classification of Cookbooks: *What's my next purchase?*

Bhanu Yerra
Metis Data Science Bootcamp, Project #4

Introduction

Ever since I started watching *Great British Baking Show*, I added three books to my ever-increasing collection of cookbooks. Do I bake? No sir, I do not! At least not very well. I am not alone in this obsession with cooking shows, and cookbooks. According to Forbes and NBCNews.com, annual sales of cookbooks have increased by 20% or over in at least the last two years [1, 2, 3]. The format of these books has changed too – more glossy paper with pictures than a collection of over-running text documenting the recipes.

So, how do we find our cookbooks? We start at Amazon.com reviews, of course! Combining my fascination with cookbooks and my passion for Data Science, I would like to help my fellow cookbook-enthusiasts in finding their next purchase by building a recommender system based on book collections and user reviews at Amazon.com. I am planning on using Natural Language Processing and unsupervised learning in clustering books together in finding topics and trends.

Data

A collection of user reviews and product information from Amazon.com is available from University of California, San Diego Computer Science department [4]. This dataset spans from 1996 to 2018, and was trimmed to include only cookbook reviews. This resulted in about 45,000 cookbooks with 29,000 (65%) of them with no user reviews, and the rest - about 16,000 cookbooks - with 428,000 reviews. This data will be further trimmed by year to include only reviews from the last 5 to 10 years.

Table 1: Data Labels for Products Table

asin	ID of the product, e.g. 0000031852
title	name of the product
feature	bullet-point format features of the product
description	description of the product
price	price in US dollars (at time of crawl)
imUrl	url of the product image
related	related products (also bought, also viewed, bought together, buy after viewing)

salesRank	sales rank information
brand	brand name
categories	list of categories the product belongs to
tech1	the first technical detail table of the product
tech2	the second technical detail table of the product
similar	similar product table

Table 2: Data Labels for Reviews Table

reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	ID of the product, e.g. 0000013714
reviewerName	name of the reviewer
vote	helpful votes of the review
style	a dictionary of the product metadata, e.g., "Format" is "Hardcover"
reviewText	text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime	time of the review (raw)

Additional Considerations

Building a recommender system based on user reviews will limit the system to only 16,000 (45%) of the books that have reviews. These might be "money makers" in the cookbook space but to maximize the sales potential a retailer like Amazon.com will be interested in building a content-based recommender system in addition to collaborative recommender system. I am planning on using the text "description" of the cookbooks in discovering clusters for all 45,000 cookbooks available on Amazon.com to build a content based recommender system.

References:

1. Cookbook Sales Are Jumping, Which Is Great News For Shops That Specialize In Them, *Forbes*, March 2019.
2. Culinary Bookstores Feed Local Appetites, *Publishers Weekly*, March 2019.
3. Recipe for success: Cookbook sales survive shift to digital media, *NBCNews.com*, August 2018.
4. Amazon Review Dataset (2018), Dr. Julian McAuley, Computer Science Department, University of California, San Diego (Data Hosted by Jianmo Li)