Capstone Approach

Brandon Sanders

May 24, 2018

# Data Science Capstone, Cohort 1

## Walmart.com Product Catalog

Walmart Labs provides an API to pull the catalog of all Walmart.com products. The data is updated every 24 hours, but delta files are available to show changes made within the last few hours. API calls are limited to 5 calls per second and 5,000 per day.

## Parent\_Category\_Title Total\_Related\_Category  
## 1 Arts, Crafts & Sewing 22  
## 2 Auto & Tires 13  
## 3 Baby 21  
## 4 Beauty 11  
## 5 Books 50  
## 6 Cell Phones 11

## Project Approach

The goal of this capstone project is to develop a process using NLP to categorize titles by their title name, short description, and long description. Challenges identified include the size of the data, variety of categories, and opportunity for a title to fall into multiple categories. The Book parent alone has 50 categories, with 1,282 subcategories.

I will use the Walmart dataset, split as training and test data, and target specific categories and sub-categories that are more easily identifiable, such as Mother’s Day, Father’s Day, Pride Month, and Summer Reading Lists. Using the category codes identified in the initial API requests, I can pull title level data for targeted categories.

## cat cat\_id sub\_cat  
## 1 Arts & Entertainment 3920\_582321 Architecture  
## 2 Arts & Entertainment 3920\_582321 Art  
## 3 Arts & Entertainment 3920\_582321 Design  
## 4 Arts & Entertainment 3920\_582321 Music Books  
## 5 Arts & Entertainment 3920\_582321 Performing Arts Books  
## 6 Arts & Entertainment 3920\_582321 Photography  
## 7 Back to School Books 3920\_8955258 Back to School Books  
## 9 Biographies & Memoirs 3920\_5196470 Adventurers & Explorers  
## sub\_id  
## 1 3920\_582321\_4942684  
## 2 3920\_582321\_6907922  
## 3 3920\_582321\_9365291  
## 4 3920\_582321\_1318217  
## 5 3920\_582321\_6572320  
## 6 3920\_582321\_5674104  
## 7 3920\_8955258  
## 9 3920\_5196470\_9289082

## itemId name  
## 1 20438802 I Am J  
## 2 15022536 Queer America : A People's Glbt History of the United States  
## 3 1996733 Same-Sex Unions in Premodern Europe

## Project Steps

1. Pull category data from API, identify categories to pull, and ingest item level data.
2. Identify necessary fields, clean data for missing values and irrelevant symbols.
3. Develop training and test datasets for developing NLP models
4. Consider mapping company sales trends with holiday title selections by category (Mother’s Day title sales annually, expect a peak in May)
5. Consider NLP visualizations, clustering, category overlap