Capstone Project: Individual Personalities with Amazon Reviews

Tadaaki Sun

Springboard

Data Science Intensive

Capstone Project Proposal: Innovating Amazon’s Recommendation System Using Sentiment Analysis

**General Overview**

Amazon’s current recommendation system uses the item-to-item collaborative filtering, which simply recommends the neighboring product of previously purchased items. As we are looking for new ways to innovate and respond to our customers need, researching valuable insights from people’s review can potentially improve our model to give a wider variety of recommendation. As we learned that human’s sometimes do not understand their actual needs, we will observe the hidden meanings of the reviews using sentimental analysis to generate scores that will identify new customer segments, which the products will be recommended based on those groups.

**General Objectives**

Improve the average rating and votes based on predictive models created via sentimental analysis scores. We will create four unique scores based on basic sentimental features (positive, negative, neutral), advanced features (increase range of positivity to negativity), subjective vs. objectiveness, and feature-based. We will use Amazon Review datasets (Fine Foods and Unlocked Cell Phones) to predict if a new reference system will recommend products for each customer segments based on the various scorings.

**Specific Objectives**

* Clean, normalize data and create additional features (word count, score based on word length, sentiment score) Connect with API - <http://text-processing.com/demo/sentiment/>
* Apply exploratory data analysis and visualization techniques to gain insight into the relationships between rating, word count, and various sentiment scores.
* Create a predictive model that can be used to improve the ratings and increase votes by various sentimental analysis
* Create a list of actionable recommendations in the form of reports and presentation slides from analysis results to the sales and marketing team of Amazon
* Create a new recommendation system with Python, and provide the code

My goal is to explore the possibilities of having sentimental analysis scores be a predictor to increasing the average scores and number of votes through identifying niche customer segments. I would investigate two different Amazon Review categories, Fine Foods and Phones where the details are in the Appendix section, which includes different levels range of products to test the legitimacy of the model to be used across all Amazon reviews.

Assuming certain values hidden in the review.

Limitation: This doesn’t factor those who did not like the product for any reasons

**Appendix: Data Overview**

Fine Foods Definition

Non-everyday food of high quality or specially prepared food, for example caviar, oysters, frog’s legs, salads or wines.

Fine Foods Data

The dataset consists of reviews of fine foods from the Amazon website: ([http://snap.stanford.edu/data/web-FineFoods.html](http://snap.stanford.edu/data/web-FineFoods.html%20/t%20_blank)).

Number of Reviews: 568,454

Number of Users: 256,059

Number of Products: 74,258

Users with >50 reviews: 260

Median words per review: 56

Timespan: Oct 1999-2012

Sample Data Format

* product/productId: B001E4KFG0
* review/userId: A3SGXH7AUHU8GW
* review/profileName: delmartian
* review/helpfulness: 1/1
* review/score: 5.0
* review/time: 1303862400
* review/summary: Good Quality Dog Food
* review/text: I have bought several of the Vitality canned dog food products and have

found them all to be of good quality. The product looks more like a stew than a

processed meat and it smells better. My Labrador is finicky and she appreciates this

product better than most.

Data Information

* product/productId: [asin](http://en.wikipedia.org/wiki/Amazon_Standard_Identification_Number), e.g. [amazon.com/dp/B001E4KFG0](http://amazon.com/dp/B001E4KFG0/)
* review/userId: id of the user, e.g. [A3SGXH7AUHU8GW](http://www.amazon.com/gp/cdp/member-reviews/A3SGXH7AUHU8GW)
* review/profileName: name of the user
* review/helpfulness: fraction of users who found the review helpful
* review/score: rating of the product
* review/time: time of the review (unix time)
* review/summary: review summary
* review/text: text of the review

Need to convert unix time code into actual time (<http://www.unixtimestamp.com/index.php>).

Unlocked Mobile Phones Data

Fine Foods Data

Unlocked Phone Data

The dataset consists of reviews of Unlocked Mobile Phones from the Amazon website: (<https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones>).

Number of Reviews: 413,840

Sample Data Format

* product name: "CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D700\*FRONT CAMERA\*ANDROID\*SLIDER\*QWERTY KEYBOARD\*TOUCH SCREEN
* brand name: Samsung
* price: 199.99
* rating: 5.0
* reviews: I feel so LUCKY to have found this used (phone to us & not used hard at all), phone on line from someone who upgraded and sold this one. My Son liked his old one that finally fell apart after 2.5+ years and didn't want an upgrade!! Thank you Seller, we really appreciate it & your honesty re: said used phone.I recommend this seller very highly & would but from them again!!
* reviews vote: 1

Data Information

* product name: name of the product
* brand name: brand of the product
* price: price of the product
* rating: rating of the product
* reviews: text of the review
* reviews vote: fraction of users who found the review helpful

Reference

https://en.wikipedia.org/wiki/Sentiment\_analysis