

Topic Modelling to Extract Information from Nutraceutical Product Reviews

Deena Liz John, Ernest Kim, Kunal Kotian, Ker Yu Ong, Tyler White, Luba Gloukhova, Diane Myung-kyung Woodbridge & Nicholas Ross

University of San Francisco

Overview

- Motivation
- ShopSmart.ml as a solution
- System Overview
- Data Source and Processing
- Topic Extraction
- Summary



UNIVERSITY OF
SAN FRANCISCO

Master of Science
in Data Science

Motivation

Motivation

- In 2016 more than 2/3 of Americans purchased dietary supplements.
- Dietary Supplements:
 - Multi-vitamins, fish-oil, single-compound chemicals (Vitamin C, etc.)
 - “Nutraceuticals”
- Incredible growth in on-line purchases and subscriptions to these products.
- Most purchases via general e-commerce platforms (Amazon, Walmart, etc.).

Sadly

- A significant downside to purchasing nutraceuticals via general e-commerce platforms is that there are over-generalized rating systems.
- Consumers:
 - Making an informed purchase is difficult.

Problem: Over-generalized rating system





Banana Republic
Banana Republic Factory Men's Crew Neck Short Sleeve Premium-Wash T Shirts
 28 customer reviews

Price: **\$12.89 - \$24.99** & Free Return on some sizes and colors

Fit: As expected (74%) ▾

Color: Charcoal



Size:

Select ▾ Size Chart

- Banana Republic
- Crew Neck Short Sleeve
- 60% cotton 40% polyester
- Soft And Comfortable Material

 Report incorrect product information.

Problem: Over-generalized rating system



Roll over image to zoom in

Le Creuset Signature Enamelled Cast-Iron 7-1/4-Quart Round French (Dutch) Oven, Cerise (Cherry Red)

by Le Creuset

★★★★★ 253 customer reviews | 86 answered questions

List Price: \$379.95

Price: \$340.00

You Save: \$39.95 (11%)

Get \$70 off instantly: Pay \$270.00 upon approval for the Amazon Prime Rewards Visa Card.

Note: Not eligible for Amazon Prime. Available with free Prime shipping from other sellers on Amazon.

Color: Cherry



- 45% larger handles that provide a sure grip, even with oven mitts
- The superior heat distribution and retention of le creuset enameled cast iron
- An advanced sand-colored interior enamel with even more resistance to wear
- A larger composite knob that withstands temperatures up to 500°F
- Colorful, long-lasting exterior enamel that resists chipping and cracking

Compare with similar items

Used & new (9) from \$338.51 ✓prime

Report incorrect product information.

Problem: Over-generalized rating system



Upper Bounce Trampoline and Enclosure Set Equipped with The Easy Assemble Feature

by [Upper Bounce](#)

 203 customer reviews | 67 answered questions

Price: **\$329.98** 

Get \$70 off instantly: Pay \$259.98 upon approval for the Amazon Prime Rewards Visa Card.

Style Name: **14-Feet**

10-Feet
\$202.99


12-Feet
\$329.99


14-Feet
\$329.98


15-Feet
from 5 sellers

16-Feet
from 13 sellers

7.5-Feet
\$174.97


Service: [Get expert assembly Details](#)

Get \$50 off assembly | Discount applied at checkout

Without expert assembly

Include installation

+\$59.75 \$109.75 per unit

 See more

- The new EASY ASSEMBLE/DISSEMBLE enclosure poles feature

- NO COMPLICATED HARDWARE NEEDED

- High Quality Trampoline Mat manufactured from premium HD mesh material

Vitamins?

Nature Made Vitamin D3 5000 IU Ultra Strength Softgels 90 Ct by [Nature Made](#)



720 customer reviews | 31 answered questions



1+



About the product

- Nutrition and Wellness
- Core Letter Vitamins
- vitamin-d
- Vitamin D helps support teeth, bone and immune health

Problem: Over-generalized rating system

- From an information theory perspective, this is a difficult problem for consumers:
 - Over a dozen fish oil pills with more than 4 stars on Amazon right now!
 - How does a consumer choose?
- All of these reviews could be legitimate – but reviewers could be focusing on different aspects of the product when making their review.

System Overview

Solution: ShopSmart.ml

- Create more informationally useful measures for consumers.
- Proof of concept using Natural Language Processing (“NLP”) techniques on nutraceutical product reviews.
- Our system provides three sets of information:
 1. Easy to read ratings for the most useful information (“Topic Score”)
 2. List of other topics that are commonly in the reviews (“Other Topics”)
 3. Full reviews which are informationally dense (“Representative Reviews”)

Information

- Topic Scores: 1-5 rating based on three important measures:
 1. Cost: How do reviewers feel about the cost?
 2. Efficacy: Does it work?
 3. Service: How well was the good provided?
- Other Topics: What "other things" are spoken of in the reviews?
- Representative Reviews: Provide a few reviews which demonstrate the above.
- Let's see how it looks!

Sample Page

ShopSmart.ML

Carlson The Very Finest Fish Oil Liquid Omega-3 Lemon, 500ml



Out of 428 reviews...



Very Finest Fish Oil Lemon Flavor by Carlson Laboratories 500 ml Liquid The Very Finest Fish Oil on salads It has a refreshing lemon taste. THIS PRODUCT is regularly tested (using AOAC internat detectable levels of mercury cadmium lead PCB's and 28 other contaminants.

[Buy it now!](#)

Topic Scores

Cost :

Efficacy :

Service :



Other Topics

- Flavor Taste : Taste, Mix, Flavor, Oil, Fish
- Product Form : Swallow, Easy, Water, Like, Small
- Purchase Related : Amazon, Price, Brand, Quality, Vitamin

Representative Amazon Reviews

Cost

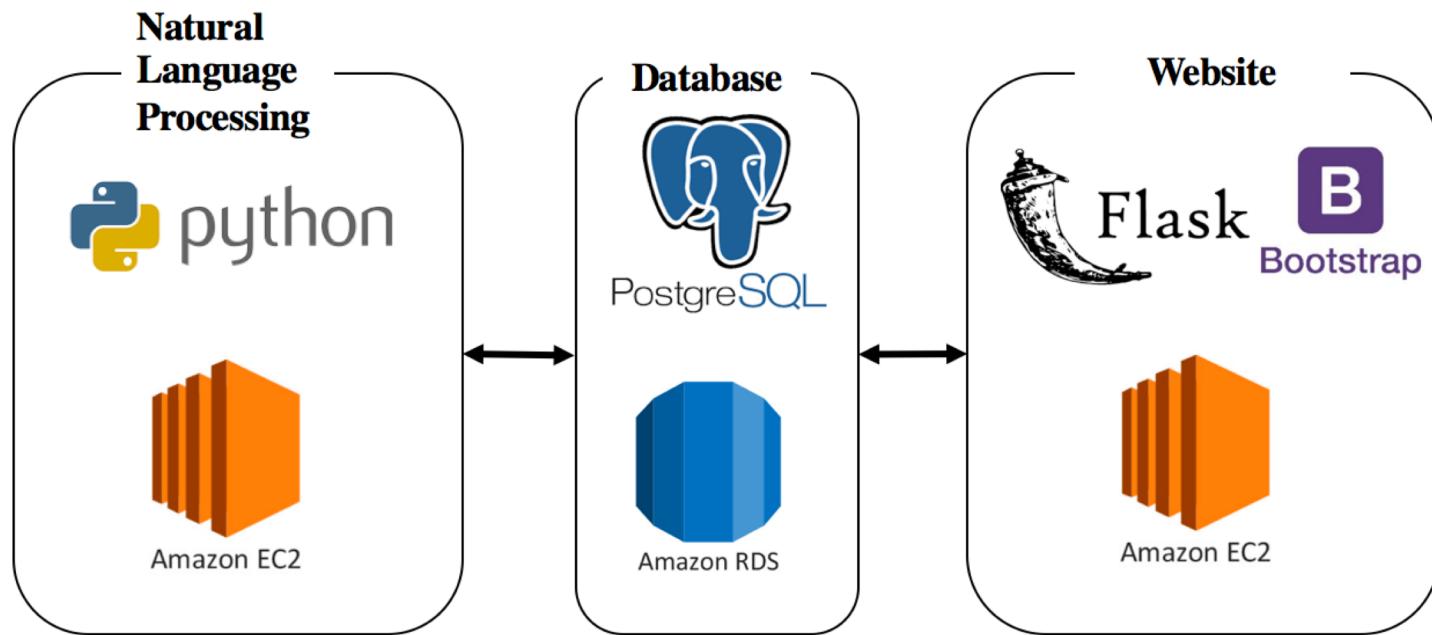
I've been using fish oil for a long time, and my children's doctor recommended it for them too. Standard capsu cheaper than Carlson. But Carlson is even higher in omega-3 than the enteric capsules, and is easy for children you. And you have to take a fist full of capsules to get the benefit of a spoonful of this, so it's far more practic gastronomically is a bit of a nice lemon taste.UPDATE: After using it for a while, I found that by putting the spi something preventable.

we just love this company. Not only is it cheapest but the shipping is so fast. it comes well packaged and I dor



Technical Systems

- System Overview



Data Processing

(the hard part)

Data Source

- Amazon Product Reviews between 1996 and 2014 (McAuley 2015).
- Contains 26,818 Single Compound products (e.g. Fish Oil/Vitamin C).

Title: “Happy Heart”

Review: “I feel better with taking these daily. My triglycerides were through the roof and since taking these, they have reduced greatly into normal levels”

Title: “Great Product and a good price through Amazon!”

Review: “Fish oil has also helped me boost my good cholesterol and I have the lab numbers to prove it. Nature Made is a good brand and buying it through Amazon is very economical.”

Data Processing – Filter Reviews

- We then filter based on the following criteria:
 1. Only 5-star reviews (over $\frac{1}{2}$ the reviews)
 2. Removed those with a title, but no text
 3. Removed those with more than 2,000 words (1% of the reviews)
 4. Remove “EDITED” reviews
- 217,530 Reviews *after* all filters are complete

Data Processing – Parsing Reviews

- For each review we did the following:
 1. Added the title as a separate sentence
 2. Tokenized (broken into a list of constituent words)
 3. De-stemmed and Lemmatized
 4. Removed URLs
 5. Removed punctuation
 6. Tagged each word/token with its part of speech

Data Processing – Phrase detection

- Once the reviews were processed we then ran a phrase detection algorithm.
- The purpose of phrase detection is to identify those words that are a single phrase and shouldn't be treated separately, such as *side effect*.
- We used Normalized Pointwise Mutual Information Score ("NPMI") which calculates, for each pair of words the following:

$$NPMI(x, y) = \left(\log \frac{p(x, y)}{p(x)p(y)} \right) / -\log p(x, y)$$

- Where $p(x)$ and $p(x,y)$ represent the number of times that the words appear in isolation or as a phrase. Anything above .6 was called a phrase.

Data Processing – Phrase detection (part 2)

- Larger values of the NPMI imply that the set of words appear as a phrase, rather than as single words.
- Before calculating NPMI we removed words like "of", "and", "with", "without" and "or".
- We then identified "bi-grams" and "tri-grams" (two and three word phrases respectively).
- Only those phrases which conformed to certain parts of speech were kept as phrases: [noun or adjective][noun] or [noun or adjective][any part-of-speech][noun or adjective] (Johnson 2006).
- Most commonly found: "side effect" "fish oil" "high quality".

Final Steps – Filtering Tokens

- After phrase detection, final step was to filter out words/tokens:
 - Very rare (< 10 reviews) tokens were removed.
 - Very common (more than 50%) were removed.
 - Single character tokens.
 - Numbers.
 - Stop words (“the”, “is”, “at”, etc.) were removed.
- After all processing there were 13,850 distinct tokens across 217,530 reviews.

Phew... data is now clean!

- So, looking back at what we did:
 - Filtered Reviews.
 - Parsed Reviews.
 - Identified phrases that should be treated as one.
 - Filtered tokens.
- This looks like a lot of arbitrary decisions, but it is all off-the-shelf, standard, NLP.
- This corresponds to a “bag-of-words” approach to NLP. We treat each review as a bag of words and then compute statistics on the distribution of words.

Topic Extraction

Topics -- Overview

- Now that our data is clean, we want to understand *what* a review is referring to.
- This is called “Topic Modelling” or “Topic extraction”
- Goal: Look for words which indicate that a review is referring to a topic.
- Start by understanding our Topic Scores (Cost, Service, Efficacy)

Cost, Efficacy and Service

- We create two word lists – one using Word2Vec and one using Anchored CorEx for the topics of “Cost”, “Efficacy” and “Service”.
- We first train a custom word2Vec model on the corpus of reviews and then look for words which are similar to “cost”, “efficacy” and “service”.
- Our second Word List is generated using an Anchored Correlation Explanation (“CorEx”) model, which uses information theory to maximize the total correlation of groups of words.

Summary

Category	Word2Vec	Anchored CorEx (additional)
Efficacy	73	9
Cost	48	8
Service	89	5

- This shows us the number of words in the word list for each topic

Returning to ShopSmart.ml

- To create the rating, we look at the percentage of reviews for a product which have a word from the word list.
- We then assign a 1 to 5 rating based on the percentage's quintile over all reviews.
- Done for Cost, Efficacy and Service

ShopSmart.ML

Carlson The Very Finest Fish Oil Liquid Omega-3 Lemon, 500ml

★★★★★

Out of 428 reviews...



Very Finest Fish Oil Lemon Flavor by Carlson Laboratories 500 ml Liquid The Very Finest Fish Oil on salads It has a refreshing lemon taste. THIS PRODUCT is regularly tested (using AOAC internat detectable levels of mercury cadmium lead PCB's and 28 other contaminants.

[Buy it now!](#)

Topic Scores

Cost : 

Efficacy : 

Service : 

Other Topics

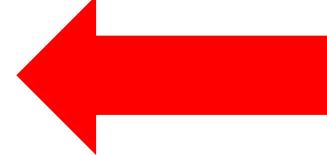
- Flavor Taste : Taste, Mix, Flavor, Oil, Fish
- Product Form : Swallow, Easy, Water, Like, Small
- Purchase Related : Amazon, Price, Brand, Quality, Vitamin

Representative Amazon Reviews

Cost

I've been using fish oil for a long time, and my children's doctor recommended it for them too. Standard capsu cheaper than Carlson. But Carlson is even higher in omega-3 than the enteric capsules, and is easy for childre you. And you have to take a fist full of capsules to get the benefit of a spoonful of this, so it's far more practic gastronomically is a bit of a nice lemon taste.UPDATE: After using it for a while, I found that by putting the spi something preventable.

we just love this company. Not only is it cheapest but the shipping is so fast. it comes well packaged and I dor



Returning to ShopSmart.ml

- For “Other Topics” we create an unanchored CorEx model which identifies groups of highly correlated words.

ShopSmart.ML

Carlson The Very Finest Fish Oil Liquid Omega-3 Lemon, 500ml



Out of 428 reviews...



Very Finest Fish Oil Lemon Flavor by Carlson Laboratories 500 ml Liquid The Very Finest Fish Oil on salads It has a refreshing lemon taste. THIS PRODUCT is regularly tested (using AOAC international) for detectable levels of mercury cadmium lead PCB's and 28 other contaminants.

[Buy it now!](#)

Topic Scores

Cost :



Efficacy :



Service :



Other Topics

- Flavor Taste : Taste, Mix, Flavor, Oil, Fish
- Product Form : Swallow, Easy, Water, Like, Small
- Purchase Related : Amazon, Price, Brand, Quality, Vitamin

Representative Amazon Reviews

Cost

I've been using fish oil for a long time, and my children's doctor recommended it for them too. Standard capsules are cheaper than Carlson. But Carlson is even higher in omega-3 than the enteric capsules, and is easy for children to swallow. And you have to take a fist full of capsules to get the benefit of a spoonful of this, so it's far more practical gastronomically. It has a nice lemon taste. UPDATE: After using it for a while, I found that by putting the capsules in a smoothie, it's easier to digest. I also like the taste.

we just love this company. Not only is it cheapest but the shipping is so fast. it comes well packaged and I don't have to worry about it getting damaged in the mail.



Unsupervised Topic Extraction

- For “Other Topics” we use Unanchored CorEx, a type of Unsupervised learning.
- The technique creates lists of words which represent a “topic”
- Using 13 topics, the model generates 8 groups of words with easily definable characteristics.
- The other five were not easily definable.

Common Ailments	Flavor / Taste
Work-out related	Chronic Ailments
Purchase-related	Appearance-related
Product form	Gut health

Returning to ShopSmart.ml

- We then look at which of the “Other Topics” are most represented in the reviews and present them.
- Presented by each is a list of the words that generated the “Other Topic”.

ShopSmart.ML

Carlson The Very Finest Fish Oil Liquid Omega-3 Lemon, 500ml



Out of 428 reviews...



Very Finest Fish Oil Lemon Flavor by Carlson Laboratories 500 ml Liquid The Very Finest Fish Oil on salads It has a refreshing lemon taste. THIS PRODUCT is regularly tested (using AOAC international) detectable levels of mercury cadmium lead PCB's and 28 other contaminants.

[Buy it now!](#)

Topic Scores

Cost :



Efficacy :



Service :



Other Topics

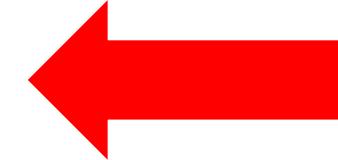
- Flavor Taste : Taste, Mix, Flavor, Oil, Fish
- Product Form : Swallow, Easy, Water, Like, Small
- Purchase Related : Amazon, Price, Brand, Quality, Vitamin

Representative Amazon Reviews

Cost

I've been using fish oil for a long time, and my children's doctor recommended it for them too. Standard capsules are cheaper than Carlson. But Carlson is even higher in omega-3 than the enteric capsules, and is easy for children to swallow. And you have to take a fist full of capsules to get the benefit of a spoonful of this, so it's far more practical gastronomically. It has a nice lemon taste. UPDATE: After using it for a while, I found that by putting the capsules in a spoonful of oil, it makes it easier to swallow. It's also something preventable.

we just love this company. Not only is it cheapest but the shipping is so fast. it comes well packaged and I don't have to worry about it getting damaged in the mail.



Returning to ShopSmart.ml

- Finally, for each of our 3 topics of Cost, Efficacy and Service we present a number of reviews which speak about each.

ShopSmart.ML

Carlson The Very Finest Fish Oil Liquid Omega-3 Lemon, 500ml



Out of 428 reviews...



Very Finest Fish Oil Lemon Flavor by Carlson Laboratories 500 ml Liquid The Very Finest Fish Oil on salads It has a refreshing lemon taste. THIS PRODUCT is regularly tested (using AOAC internat detectable levels of mercury cadmium lead PCB's and 28 other contaminants.

[Buy it now!](#)

Topic Scores

Cost :



Efficacy :



Service :



Other Topics

- Flavor Taste : Taste, Mix, Flavor, Oil, Fish
- Product Form : Swallow, Easy, Water, Like, Small
- Purchase Related : Amazon, Price, Brand, Quality, Vitamin

Representative Amazon Reviews

Cost

I've been using fish oil for a long time, and my children's doctor recommended it for them too. Standard capsu cheaper than Carlson. But Carlson is even higher in omega-3 than the enteric capsules, and is easy for children you. And you have to take a fist full of capsules to get the benefit of a spoonful of this, so it's far more practic gastronomically is a bit of a nice lemon taste.UPDATE: After using it for a while, I found that by putting the spi something preventable.

we just love this company. Not only is it cheapest but the shipping is so fast. it comes well packaged and I dor



Summary

Summary

- Nutraceuticals are a growing business, but purchasing them online is difficult.
- Information presented by general e-commerce platforms fails to cater to consumers in this space.
- Using customer reviews we have built a proof of concept which allows consumers to more efficiently make decisions.

Thanks!

Questions: ncross@usfca.edu