Identifying Customer Needs from Amazon Reviews

Artem Timoshenko <atimoshe@mit.edu>
Github: https://github.mit.edu/atimoshe/Customer-Needs

Introduction

Understanding customer needs is essential for successful new product development. The traditional approach to identifying customer needs is interview-based and requires manual analysis, which is expensive. Nowadays, millions of customers describe their preferences for products at online reviews, but marketing managers rarely explore online content, as it is noisy and often repetitive. At the current project, I am trying to identify a set of non-repetitive customer needs from Amazon reviews for a particular product category, and show how to organize them into a convenient structure.

Data

I use the Amazon product reviews dataset[1], which contains over 140 million product reviews and metadata spanning May 1996 – July 2014. At this project, I concentrate on the 'Oral Care' product category, and operate with ~138'000 reviews.

I also have data from a marketing consulting agency, which regularly conducts a customer needs analysis. The agency provided me with transcripts of interviews for several product categories, including oral care products, and final hierarchies of needs constructed based on these interviews. They have also provided me with training materials for customer needs identification.

Unfortunately, interview transcripts from the agency could not be used as a training data, as content and language are very different at the interviews and online reviews. However, I use training materials to do my own labeling. In particular, I randomly select ~1'000 sentences from Amazon review dataset, and label them into 'informative' and 'non-informative'. Then I randomly generate ~200 pairs of 'informative' sentences and label distances between them as 0/1/2/3 to identify how close these sentences should appear at the final hierarchy: 0 is for completely identical sentences, 1 corresponds to the same secondary need level, 2 is for the same primary need category, and 3 is for completely unrelated customer needs.

[1] http://jmcauley.ucsd.edu/data/amazon/

Solution Strategy

There are two primary goals of this project: (1) identify a list of ~500 non-repetitive customer needs for a given product category, and (2) organize the list of customer needs into a hierarchy[2]. Here is an outline of the approach I use:

0. Split reviews into sentences + Sentence preprocessing

   I split Amazon reviews for Oral Care category into ~600'000 sentences using NLTK, and do basic sentence preprocessing to make them consistent with Google News 300-dim word2vec. In particular, I transform sentences into lowercase, drop stop words, drop punctuation and special symbols except % and $, change numbers with # (12-> ##, 5-> #), etc.

1. Identify 'informative' sentences

   About 80% sentences at the reviews contain completely no information about customer needs. Thus, I classify sentences into 'informative' and 'non-informative' before proceeding to further analysis. Classification is based on 1'000 sentences, which were randomly drawn from the dataset. I found out later that even simple sentence embeddings are able to catch similarities between sentences, which could be used to organize a smarter labeling procedure, but at this stage of the project, the training set is very unbalanced (over 80% zeros).

   I do a logistic regression on sentence embeddings as a baseline for sentence classification. Sentence embeddings here are calculated as an average of word2vec representations of words in the sentence excluding stop-words. A more advanced approach is to train a convolutional neural network, and I use an implementation by Tao Lei, Regina Barzilay, and Tommi Jaakkola (2015).

2. Remove duplicates + Build a hierarchy

   Convolutional neural network identifies ~40'000 informative sentences at the dataset. A lot of these sentences are close in their meaning, so I want to remove duplicates, and I expect to get ~500 non-repetitive sentences. Moreover, having a list of 500 items is not convenient for further analysis, so I want to organize them into a convenient structure.

[2] See example at Appendix A

A key problem here is how to define a distance between two sentences. I try three different approaches: (1) standard cosine or Euclidean distances between sentence embeddings, (2) supervised prediction based on the difference of sentence embeddings: $(x_i - x_j)M(x_i - x_j) \in \mathbb{R}^+$, and (3) supervised prediction of the general form: $f:(x_1, x_2) \to \mathbb{R}^+$, where $f(\cdot,\cdot)$ is a symmetric function. I also experiment with different types of sentence embeddings: (1) simple or tf-idf weighted average of word2vec embeddings, (2) softmax input at the CNN trained at part 1, and (3) spectral embeddings based on (1) and (2).

Results

Part 1 – Identify informative sentences

	Precision	Recall	F1 score
LR-SA	0.367	0.440	0.400
LR-WR	0.356	**0.525**	0.425
CNN-L	**0.680**	0.315	**0.430**
CNN-C	0.667	0.222	0.333

Measures: (1) Precision = TP / (TP+FP); (2) Recall = TP / (TP+FN);

(3) F1 = 2*Precision*Recall / (Precision + Recall)

Methods: (1) LR-SA = Logistic regression; Sentence embeddings: simple average of word2vec

(2) LR-WR = Logistic regression; Sentence embeddings: tf-idf weighted average of word2vec

(3) CNN-L = Convolutional Neural Network; Softmax input: last layer output

(4) CNN-C = Convolutional Neural Network; Softmax input: concatenation of outputs of all layers

I use only 500 sentences for training, but both baseline and CNN approaches show quite high precision and recall. Taking into account that the training set is very unbalanced, I find these results promising. Also note that CNN slightly outperforms baseline even with a small training sample.

Part 2 – Build a hierarchy

I have experimented with multiple techniques to define a distance function and to calculate sentence embeddings, as described at the Solution Strategy part. An important

problem here is how to compare different approaches. I tried to use 200 labeled pairs of sentences to build distance predictions at the supervised fashion, and evaluated the average of squared errors. Supervised methods produced meaningless results, probably, because the training set was too small.

I ended up having the following informal evaluation procedure. I pick up 9 sentences: 3 of them about white teeth, 3 are about fresh breath, and 3 are about convenient packaging[3]. Ideally, we should be able to cluster these 9 sentences into 3 groups using a distance function. Having well-defined clusters will allow removing too similar sentences. To build a hierarchy of identified needs, we want to have one more property of the distance function: clusters for white teeth and fresh breath should be closer to each other than teeth-packaging or breath-packaging. So, the ideal distance matrix for these 9 sentences should look as follows:

	0	1	2	3	4	5	6	7	8
0	0.000	0.250	0.250	0.500	0.500	0.500	0.700	0.700	0.700
1	0.250	0.000	0.250	0.500	0.500	0.500	0.700	0.700	0.700
2	0.250	0.250	0.000	0.500	0.500	0.500	0.700	0.700	0.700
3	0.500	0.500	0.500	0.000	0.250	0.250	0.780	0.780	0.780
4	0.500	0.500	0.500	0.250	0.000	0.250	0.780	0.780	0.780
5	0.500	0.500	0.500	0.250	0.250	0.000	0.780	0.780	0.780
6	0.700	0.700	0.700	0.780	0.780	0.780	0.000	0.250	0.250
7	0.700	0.700	0.700	0.780	0.780	0.780	0.250	0.000	0.250
8	0.700	0.700	0.700	0.780	0.780	0.780	0.250	0.250	0.000

Here are my best results:

	0	1	2	3	4	5	6	7	8
0	0.000	0.525	0.125	1.006	1.083	1.063	0.969	1.101	1.035
1	0.525	0.000	0.551	1.160	0.900	0.912	1.026	1.068	1.037
2	0.125	0.551	0.000	0.949	1.018	1.033	1.056	0.996	1.003
3	1.006	1.160	0.949	0.000	0.721	0.912	1.115	0.990	1.057
4	1.083	0.900	1.018	0.721	0.000	0.691	1.252	1.104	1.072
5	1.063	0.912	1.033	0.912	0.691	0.000	0.998	1.264	1.095
6	0.969	1.026	1.056	1.115	1.252	0.998	0.000	0.733	0.735
7	1.101	1.068	0.996	0.990	1.104	1.264	0.733	0.000	0.288
8	1.035	1.037	1.003	1.057	1.072	1.095	0.735	0.288	0.000

	0	1	2	3	4	5	6	7	8
0	0.000	0.322	0.304	0.419	0.522	0.591	0.650	0.622	0.693
1	0.322	0.000	0.255	0.399	0.476	0.493	0.643	0.654	0.715
2	0.304	0.255	0.000	0.420	0.534	0.519	0.664	0.658	0.744
3	0.419	0.399	0.420	0.000	0.311	0.525	0.604	0.535	0.517
4	0.522	0.476	0.534	0.311	0.000	0.433	0.593	0.507	0.615
5	0.591	0.493	0.519	0.525	0.433	0.000	0.684	0.694	0.787
6	0.650	0.643	0.664	0.604	0.593	0.684	0.000	0.407	0.469
7	0.622	0.654	0.658	0.535	0.507	0.694	0.407	0.000	0.332
8	0.693	0.715	0.744	0.517	0.615	0.787	0.469	0.332	0.000

Matrices at both pictures were obtained for the standard cosine distance metric. At the picture on the right, the simple average of word2vec was used as sentence embeddings. To get the picture on the left, I transformed the same embeddings into a 50-dimensional Laplacian Eigenmaps. Non-standard (trained) distance functions and CNN-based embeddings generate less clear or completely meaningless results.

[3] See the list of sentences at Appendix B

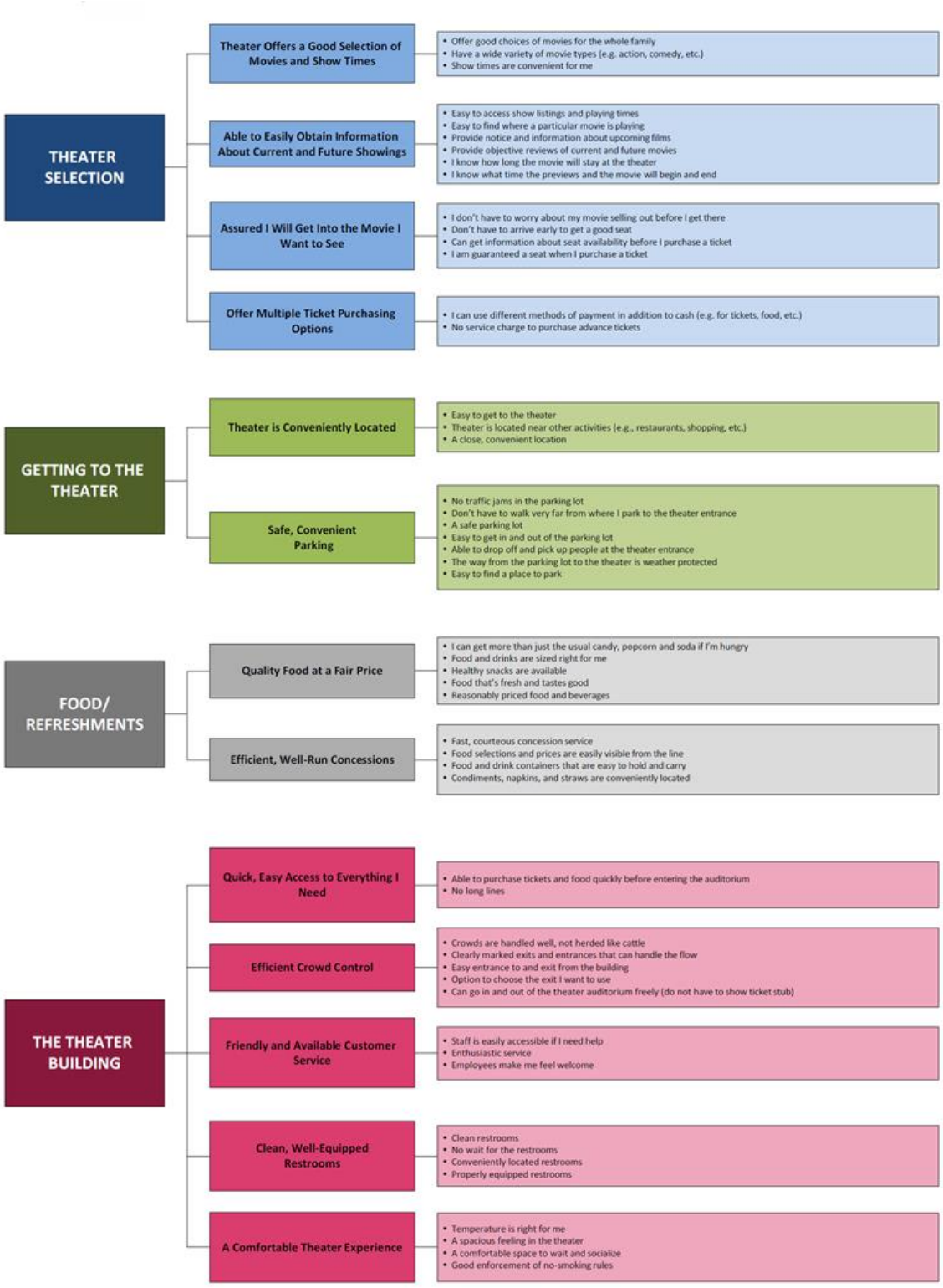
We can see that similar sentences 0-1-2, 3-4-5, and 6-7-8 cluster together at both pictures. These clusters are especially clear at the picture on the left, so the corresponding method can be used for removing duplicate sentences. On the other hand, clusters for white teeth and fresh breath stay closer to each other than teeth-packaging and breath-packaging at the picture on the right, so the corresponding method is promising for building a hierarchy.

Future Work

I have developed a basic approach for solving the problem of identifying customer needs from online reviews. It could be improved in multiple ways: (1) Smarter labeling at Part 1 to create a more balanced dataset; (2) Better of sentence embedding (doc2vec, Laplacian eigenmaps, retrofitting, etc.); (3) Different classification techniques at Part 1 (I tried only Logistic Regression and CNN. Neural nets require more labeled data, but SVM could produce good results); (4) Formal evaluation procedures for Part 2 (It is especially important for finding dimensionality of spectral embeddings).

From the social sciences perspective, a number of questions could be studied using this tool. For example, it is a systematic way of extracting customer needs from online content. It is interesting how the outputs of this algorithm compare to the results of a traditional interview-based approach. This could be a good demonstration of ‘knowledge’ contained at the online reviews.

Appendix A


THEATER SELECTION
Theater Offers a Good Selection of Movies and Show Times
• Offer good choices of movies for the whole family
• Have a wide variety of movie types (e.g. action, comedy, etc.)
• Show times are convenient for me
Able to Easily Obtain Information About Current and Future Showings
• Easy to access show listings and playing times
• Easy to find where a particular movie is playing
• Provide notice and information about upcoming films
• Provide objective reviews of current and future movies
• I know how long the movie will stay at the theater
• I know what time the previews and the movie will begin and end
Assured I Will Get Into the Movie I Want to See
• I don't have to worry about my movie selling out before I get there
• Don't have to arrive early to get a good seat
• Can get information about seat availability before I purchase a ticket
• I am guaranteed a seat when I purchase a ticket
Offer Multiple Ticket Purchasing Options
• I can use different methods of payment in addition to cash (e.g. for tickets, food, etc.)
• No service charge to purchase advance tickets
GETTING TO THE THEATER
Theater is Conveniently Located
• Easy to get to the theater
• Theater is located near other activities (e.g., restaurants, shopping, etc.)
• A close, convenient location
Safe, Convenient Parking
• No traffic jams in the parking lot
• Don't have to walk very far from where I park to the theater entrance
• A safe parking lot
• Easy to get in and out of the parking lot
• Able to drop off and pick up people at the theater entrance
• The way from the parking lot to the theater is weather protected
• Easy to find a place to park
FOOD/ REFRESHMENTS
Quality Food at a Fair Price
• I can get more than just the usual candy, popcorn and soda if I'm hungry
• Food and drinks are sized right for me
• Healthy snacks are available
• Food that's fresh and tastes good
• Reasonably priced food and beverages
Efficient, Well-Run Concessions
• Fast, courteous concession service
• Food selections and prices are easily visible from the line
• Food and drink containers that are easy to hold and carry
• Condiments, napkins, and straws are conveniently located
THE THEATER BUILDING
Quick, Easy Access to Everything I Need
• Able to purchase tickets and food quickly before entering the auditorium
• No long lines
Efficient Crowd Control
• Crowds are handled well, not herded like cattle
• Clearly marked exits and entrances that can handle the flow
• Easy entrance to and exit from the building
• Option to choose the exit I want to use
• Can go in and out of the theater auditorium freely (do not have to show ticket stub)
Friendly and Available Customer Service
• Staff is easily accessible if I need help
• Enthusiastic service
• Employees make me feel welcome
Clean, Well-Equipped Restrooms
• Clean restrooms
• No wait for the restrooms
• Conveniently located restrooms
• Properly equipped restrooms
A Comfortable Theater Experience
• Temperature is right for me
• A spacious feeling in the theater
• A comfortable space to wait and socialize
• Good enforcement of no-smoking rules

Appendix B

White Teeth:

- And where the tartar was, shiny white teeth showed through. And I am still using it and much of the thick tartar is gone!
- I'll just smile at her with my clean, white teeth.
- I was resigned to having dull, off-white teeth and now I have white teeth and gums that don't bleed. The two-minute timer especially helped me.

Fresh Breath:

- Bad breath is no longer a problem, either, now that our gums are healthier. In addition, the unit has a sleek, modern design that looks good in the bathroom.
- For some reason my husband was having a real problem with bad breath.
- Once applied, I sighed a breath of glorious mouth relief!

Travel Convenience:

- When you travel it closes up well.
- The travel reservoir that comes with the unit does, however, have a cover.
- Also have a portable unit with which I travel.