From the
RETAIL TECH BULLETIN
Fourth Quarter 2019

Enhancing a Retailer's Online Search Returns Using Machine Learning

By Yunxuan Wang, Tianhua Zhu, Tianyu Li, Yiqing Li, Yi Wu, Graduate Students, Integrated Marketing Communications Program, Medill, Northwestern University

From April to June 2019, a team of five graduate students at Northwestern University conducted a research project to consider ways to enhance the online user search experience for a major retailer's website. The project showcased applications of a variety of programming and machine learning techniques, including tf-idf, text-similarity, Latent Dirichlet Allocation, and K Nearest Neighbors.

At the beginning of the project, the team discovered two opportunities for search improvement on the website. One was that some search terms returned no search results. For example, if the search term entered was "Christmas gift for girlfriend" on the website, no product would be returned. However, this is a search term that a customer would very possibly use. The second issue discovered was that some customers perceive a product in a way that is different from how the website management team does. For example, in customer reviews, an item labeled by the retailer as a "casual dress" was described by a customer as a "fancy dress." This gap creates a potential challenge for generating matching search results and sales.

Only if the (search bar) results were relevant and accurate would the customer get to the last step of a purchase decision. Otherwise, the customer may leave the site, and there would not be a sales conversion.

The two issues both led to undesirable search results that were either irrelevant or null. Only if the results were relevant and accurate would the customer get to the last step of a purchase decision. Otherwise, the customer may leave the site, and there would not be a sales conversion. Working with the Retail Analytics Council AI Lab at Northwestern University, the retailer sought to resolve this discrepancy to optimize the conversion rate on its e-commerce site and achieve greater sales lift.

The team started with the premise that a model taking into account not only the objective product descriptions but also the customers' perception of the products would be able to produce more satisfying search results. Therefore, the goal is to build a machine learning model that is better at capturing the intended meaning of the search terms by incorporating customer reviews. To accomplish this, the very first step is to prepare text-data for machine learning models. The performance of the final model relies on the quality of text pre-processing.

To convert text to a format that the "machine" can understand, the team first parsed the text to remove punctuation and stop words and conducted lemmatization. After

these initial cleanings, each product had a pool of words associated with it, which was referred to as a document. Then, the team experimented with two ways of text vectorization to flag the features of each product, defined by word occurrences. The first was a simple occurrence encoding, using CountVectorizer from Python's Scikit-learn to tokenize, build a corpus, and then encode a document. The encoded vector contains an integer count of the number of times each word appears in a document. The problem with word count, however, is that commonly occurring words have large counts in the documents but provide little meaning. To address the issue, the team implemented the TF-IDF feature generation approach. Essentially, TF-IDF takes into account word occurrence both locally within a document and globally across the documents, highlighting words that are more interesting to a specific document.

With all the documents encoded, the next step is to quantify the similarities among documents and the search terms to be entered by the customers. The team tried an unsupervised version of the nearest neighbor model to find the closest instances in term of the inter-document distances represented in a vector space. To avoid Euclidean distance's disadvantage of dealing with documents of uneven lengths, the team used cosine-similarity to find the nearest samples with features named in the search terms.

This model was able to match a search term like "great Christmas gift for girlfriend" with products related to festive occasions or a gifting purpose or girlfriend, thanks to previous customer reviews that mentioned customers' post-purchase interaction with the products. This test case demonstrates that even when the customer did not even specify the desirable kind of products but specifically pointed out the occasion and purpose of the purchase, the model in training was able to provide some relevant options for further review.

However, this model still has its limit in processing the information by each individual word, ignoring both the link among words and the words not included in the current corpus. In order to capture more of the ambiguity in search terms, the team further explored topic modeling, which is a technique that helps extract hidden topics from texts. It would help identify key factors pertaining to customer online shopping experiences so that the team could make recommendations on search and non-search improvements.

The team chose to experiment with Latent Dirichlet Allocation, a topic modeling technique that is often useful for search engines, news article analysis, etc. LDA assumes that each document is generated from a collection of topics and each topic is generated from a collection of words. Given a set of documents, LDA would reverse engineer the process to find the topics that make up the documents in the first place. Using the Gensim package from Python, the team was able to implement LDA and extract several segregated and meaningful topics that unveil customer sentiment, preferences, and concerns.[1]

Overall, the team was able to identify search and non-search related solutions to improve the online shopping experience of the customers. Notwithstanding, this research has certain limitations, such as only using a few product categories to build the models, only using website data over a short time period, and lack of objective and systematic ways to test the models. However, it provided a simple demonstration of using machine learni ng techniques to solve problems in an online retail scenario.

[1] Sometimes the topic keyword may not provide enough information to make sense of a specific topic. To address the issue of insufficient information provided by the weightage of keywords in each topic, the team pulls out and examines the most important documents for each topic. By manually going over the customer reviews, the team is able to capture the important nuances between the topics and develop a more well-rounded understanding of factors salient to consumer online experience.