IST 652 Final Project Proposal

Group of: Zequn Che, Jinwen Song

**Review Texts Analysis for Amazon**

For the fast-developing E-commerce market, sellers and companies need to understand the thoughts and demands of the customers. Comprehension of the thoughts from the customers helps to enhance the quality of service. [Research](https://www.brightlocal.com/research/local-consumer-review-survey/?SSAID=314743&SSCID=31k4_jo2t) shows that 91 percent of people regularly read online reviews, and 84 percent trust online reviews as personal recommendations. There is about 93 percent of consumers spend more than a minute reading reviews. Given the above facts, there is an urging need to analyze the product reviews to improve the online business.

One of the best ways to know the thoughts of the customers is understanding the review texts of the products. We can clearly know the advantages and imperfections of certain products. We could also know what do customers care the most about products.

By analyzing the review texts of customers, sellers can improve their products to change the customers’ buying decisions. Analyzing review texts could help increasing selling and to enhance brand awareness and much more.

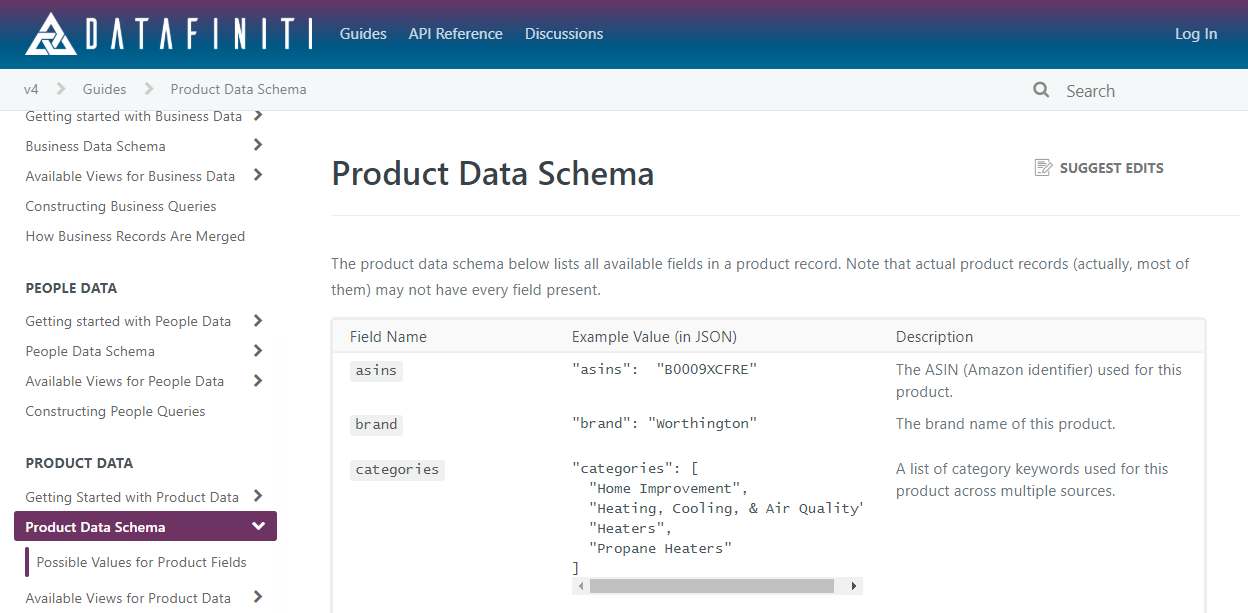
**The datasets of the plan using**

The data of the amazon review texts for analysis could be downloaded from the internet.

1. Amazon product data from <http://jmcauley.ucsd.edu/data/amazon/>



1. Amazon product data from Kaggle <https://developer.datafiniti.co/docs/product-data-schema>



**Methods we may use**

First of all, we need to break all the review texts of a certain product into words. Then we can make the words as tokens which are better for analysis.

1. The first method we can use is to count the words of a high occurrence rate by using [the countVectorizer from the Scikit-learn library](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html?highlight=countvectorizer#sklearn.feature_extraction.text.CountVectorizer). In this method, we need to remove unnecessary tokens in the word list. The punctuations and stop words that are not very valuable in this analysis will be removed. Finally, we can see the significant keywords in the big set of text data.
2. The second method is similar to the first one but mainly for understanding the basic directions of the review texts. We can analyze if a word is positive or negative then calculate the positive(or negative) rate of a product. Finally, we can know the basic customers’ attitudes toward the product.

**Potential development tasks which need additional guidance**

1. The processing work of review texts may need extra skills in language processing. We need to learn some extra skills in **language processing**.
2. To **select good data** for analysis, we need to find and read more datasets of product review texts.
3. According to the second method, when analyzing the review, we may create some word clouds. Look at occurrences of bigrams or trigrams, since “like” would also come from “doesn’t like”.