Mining Amazon Product Reviews using NLP

(Springboard Data Science career track: capstone2)

(Reeshma K)

Problem Statement: Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser. Consumers find a product of interest by visiting the website of the retailer directly or by searching among alternative vendors using a shopping search engine, which displays the same product's availability and pricing at different e-retailers. But online shopping comes with its own limitations. One of the biggest challenges is verifying the authenticity of a product. Is it as good as advertised on the e-commerce site? Will the product last more than a year? Are the reviews given by other customers true or are they false advertising? These are important questions customers need to ask before spending their money.

Ratings alone do not give a complete picture of the products we wish to purchase. So secondary option is looking at the reviews. Review's plays an important role in the decision-making process. If the number of reviews is less, it is easy to read and understand but what if there are thousands of reviews. So, the problem is How we can analyze great number of online reviews using Natural Language Processing (NLP)?

This project will serve three purposes

1. Helps consumers to understand the sentiment of the review.
2. Help the consumers to get consumer feedback in the form of topics covered by the reviews without having to go through all of them.
3. Enable consumers to quickly extract the summary of the reviews without reading the entirely.

Client: Amazon customers

Dataset: https://www.kaggle.com/snap/amazon-fine-food-reviews

Approach: The entire project is divided into three parts

Part1: Sentiment Analysis

Sentiment analysis is an important concept. Reading the sentiment of consumers will help us easily get an idea about the product. I am using TextBlob for sentiment analysis of Amazon product reviews. TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks.

Part 2: Topic Modeling

Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. I am using LDA (Latent Dirichlet Allocation) for topic modeling. LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topics.

Part 3: Text Summarization

Customer reviews can often be long and descriptive. Analyzing these reviews manually is a tedious task. Natural Language Processing can be applied to generate a summary for long reviews.