Capstone project 1: Project Proposal

1. Title: Pet Product Auto-Subcategorization by Review Analysis

2. Problem:

E-commerce companies set up categories for their products; for example, Clothing, Beauty, Books, Pet Supplies, etc. However, if the number of products in a category has been growing, they might want to classify the products into subcategories for several reasons. Because the number of products would be large, it could be difficult to categorize them one by one.

Nowadays, most e-commerce websites have reviews, which are written by customers and which help future customers decide whether they buy. On the other hand, reviews also provide much information to an e-commerce company; what customers liked or disliked, what they wanted, how they used it, and so on. In other words, reviews include information to categorize a product into subcategories.

Here, I chose Pet Supplies as a category reclassified into subcategories. I am going to make a system that automatically classifies products in Pet Supplies category into subcategories by analyzing the reviews.

3. Client:

The first clients will be e-commerce companies that would like to use reviews to subcategorize their products. Their purposes could be to improve analysis of trends and customer needs to a specific field, and/or to increase customer satisfaction by easy access to a product they want.

4. Dataset:

The animal product review data will be acquired from AWS. This data contains information about the marketplace, product ID, product name, product category, star rating, review, date, etc. In this project, I will focus on the data collected during 2014-2015 in the US.

- Amazon Pet Supplies reviews in the US (gz file) from AWS (1995 - 2015)
  https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Pet_Products_v1_00.tsv.gz

5. Approach:

The first step will be to clean and transform the data. The data which will be used to analyze the reviews will be extracted: product_parent, product_title, review_body and review_date in 2014 and 2015. Also, this step includes identifying missing data, duplicate data, and outliers.

The second step will be converting text data to numerical values. This step allows Python to analyze the review data.

Then, the products will be grouped by unsupervised clustering. Some clustering algorithms might be tried: k-mean, hierarchical clustering or DBSCAN. Once some models have been gotten, they will be tested and accomplished about at least one model.

6. Deliverables:

a. Codes (Jupyter Notebook)
b. Report
c. Presentation