**Data 620 Final Project Proposal**

**Overview:**

The success and failure of a product sold in amazon can be determined by the customer feedback which helps new customers to gain insight about the product. To minimize the time to read tons of product reviews and at the same time provide the overall summary of reviews, online retailers has to highly rely on recommendation engine. We would like to analyze on the sentiments associated with specific category “Health and Personal Care” of the products sold in amazon.

The objective is to

* Find a classification model that works best with the data.
* Understanding the topics and words that describe the broad categories of “Health and Personal Care” products sold over Amazon.

**Sources:**

http://jmcauley.ucsd.edu/data/amazon/

**Data set:** <http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Health_and_Personal_Care_5.json.gz>

Data set has the following fields:

* reviewerID - ID of the reviewer, e.g. [A2SUAM1J3GNN3B](http://www.amazon.com/gp/cdp/member-reviews/A2SUAM1J3GNN3B)
* asin - ID of the product, e.g. [0000013714](http://www.amazon.com/dp/0000013714)
* reviewerName - name of the reviewer
* helpful - helpfulness rating of the review, e.g. 2/3
* reviewText - text of the review
* overall - rating of the product
* summary - summary of the review
* unixReviewTime - time of the review (unix time)
* reviewTime - time of the review (raw)

**Approach:**

* Data Exploration:

Analyze the data set.

* Preprocessing:

Data cleanup like removing punctuations, stop words etc.

* Model data
  1. Classification / Sentiment Analysis
     + Logistic Regression
     + Multinomial Naive Bayes
  2. Clustering / Topic Modeling
     + Nonnegative Matrix Factorization (NMF)
     + Latent Dirichlet Allocation (Lda)
* Summarize Findings and Proposed Further Work

**Team:**

Dilip Ganesan, Ahmed Sajjad and Raghu Ramnath