Abstract

Anak Agung Ngurah Bagus Trihatmaja, Anupam Gupta, Ayush Bandari, Xiaomeng Peng

We will implement various unsupervised methods on Amazon Fine Food Reviews [1]. The dataset consists of reviews of fine foods from Amazon. The overview of the dataset is as at the end of this document.

In this project, we want to cluster the dataset into two clusters: positive review and negative review. To do that we will do dimensionality reduction using Principal Component Analysis, t-sne and Latent Dirichlet Allocation, and compare the results to know which method works better for this case and discuss why it works or not. For the clustering, we will try various clustering methods we have learnt in class. For example: k-means, GMM, or DBScan. Again, we will discuss the result to know which clustering methods work best for this kind of data. For the data pre-processing on the review dataset, we will follow the standard procedure on text, which are stopwords removal and stemming and then convert it into vectors of tf-idf or bag of words.

Furthermore on this analysis, we will not only utilize the reviews from the user, but also we will use a transaction data to see what combinations of products a user usually buy using FP-Growth algorithm.

[1] http://snap.stanford.edu/data/web-FineFoods.html

|  |  |
| --- | --- |
| Dataset statistics | |
| Number of reviews | 568,454 |
| Number of users | 256,059 |
| Number of products | 74,258 |
| Users with > 50 reviews | 260 |
| Median no. of words per review | 56 |
| Timespan | Oct 1999 - Oct 2012 |

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Id | Row Id |
| Unique identifier for the product | Unique identifier for the product |
| UserId | Unique identifier for the product |
| ProfileName | Profile name of the user |
| Helpfulness Numerator | Number of users who found the review helpful |
| Helpfulness Denominator | Number of users who indicated whether they found the review helpful |
| Score | Rating between 1 and 5 |
| Time | Timestamp for the review |
| Summary | Brief summary of the review |
| Text | Text of the review |