# Problem Description and Data Acquisition

Tuan Dinh
dinh5@wisc.edu

Sam Gelman
sgelman2@wisc.edu

Varun Sah
varun.sah@wisc.edu

February 8, 2017

## 1 PROBLEM DESCRIPTION

The objective of the project is to compare and contrast product listings as well as customer characteristics, preferences and bias between a specialized, niche seller of computer hardware and software (Newegg) and the electronics segment of an e-commerce giant (Amazon).

### 1.1 QUESTIONS

Some questions we hope to answer are listed below, grouped by category.

- Analysis of similarities and distinctions between Amazon and Newegg
  - Product Listings
    - Is there a significant preference for or bias against certain products or brands on either website?
    - Are product prices, on average, lower on one website?
    - Is there a correlation between a product's price and its average rating?
    - For certain products listed on both websites, is there a difference in the density distribution of customer ratings?
  - Customer characteristics
    - Are customers on one website happier than customers on the other website (sentiment analysis)?

    - Is there a significant difference in the vocabulary of customers on both websites?
    - Are customers more likely to leave longer reviews for expensive products or for cheaper ones?
    - Are more customers likely to write a review for expensive products or for cheaper ones?

- Analysis of Amazon data
  - Can we identify fake or paid reviews?
  - Can we predict if customers will find a review helpful?

# 2 Data Acquisition

## 2.1 Data Sources

Initially, we had decided to use the Amazon Electronics review data set provided by UCSD. However, we revisited our decision when we realized that the dataset did not have any reviews posted after July 2014. Also, certain product information that we desired was absent from the dataset, which meant we would have to crawl the website to get the missing information. In light of these factors, we decided to switch directions by crawling the website to extract both the products and reviews in order to have more current and updated information.

The Amazon Product Advertising API for Amazon affiliates was another option that we explored. We later decided not to utilize this option due to the usage restrictions in terms of the limited size of the API response as well as the limited number of API calls available without a commercial licence. The absence of bulk APIs meant it would be impossible to collect data of any significant proportions using this option without a commercial license.

For Newegg, deciding on the source was significantly easier. In the absence of any existing dataset or API for Newegg, crawling the website was the only alternative.

Due to the aforementioned, reasons, the respective websites of Amazon and Newegg have come to be the sources of our data:

- Amazon: http://www.amazon.com
- Newegg: http://www.newegg.com

## 2.2 METHODOLOGY

The following sections describe the process of data acquisition thus far as well as plans for the next phase of data extraction from the collected text documents.

### 2.2.1 AMAZON

We wrote scripts in Python that scraped product and review data from Amazon's website. The extracted raw data obtained by crawling is restructred and stored in the JSON format. Some of the collected data has been broken down and stored in smaller files for readability and ease of browsing. These files will later be merged into one single file before the analysis phase. The data extracted from Amazon can be found at:

- Amazon data: http://pages.cs.wisc.edu/ varun/data-science/files/Amazon/

### 2.2.2 NEWEGG

We wrote a script in Python, using the ScraPy framework, that scraped product and review data from Newegg's website. The extracted raw data obtained by crawling is restructred and stored in the CSV format. Some of the collected data has been broken down and stored in smaller files for readability and ease of browsing. These files will later be merged into one single file before the analysis phase. The data extracted from Newegg can be found at:

- Newegg data:http://pages.cs.wisc.edu/ varun/data-science/files/Newegg/

### 2.2.3 TEXT DOCUMENTS

The product descriptions and customer reviews collected by our crawlers from both Amazon and Newegg form the content of our text documents. The text documents can be found at:

- Text Documents:http://pages.cs.wisc.edu/ varun/data-science/files/TextDocuments/

DATA EXTRACTION PLAN The entity types that can be extracted from the text documents would include:

- Brand Names
- Product Names
- Adjectives used to describe the product.

Apart from identifying entity types as part of this phase, we could potentially extract sentiment from the given documents as well.

#### 2.2.4 TOOLS USED

The product descriptions and customer reviews collected by our crawlers from both Amazon and Newegg form the content of our text documents.

- ScraPy: An open source and collaborative framework for extracting data from websites. It is an application framework for writing web spiders that crawl web sites and extract data from them. We used Scrapy for extracting data from Newegg.
- Amazon Product Advertising API: Amazon's affiliate API to get product details. Was explored NOT used finally.
- Rmazon : R code to fetch amazon product and review data. Was explored but NOT used finally.