UNIVERSITY OF WATERLOO
Faculty of Mathematics

GIDE MANUAL: AN OVERVIEW OF WEB SCRAPING, PRODUCT ANALYSIS, AND REPORT OUTPUTS

Anshuk Chhibber
Waterloo, Ontario

Prepared by
Michael Brock Li
CTO Gide
ID 000000
December 21, 2018

Table of Contents

List of Tables and Figures

1.0 Introduction

Reviews started a long time ago where they are usually distributed by word of mouth or a messenger running in the streets, yelling that there is a place opening. There was no Internet; users would go to stores and buy their favourite merchandises.

The past is not a place for people nowadays; lives have become a lot easier. As technology advances throughout the years since the invention of world wide web, many people are gaining onboard of looking at reviews online. They tend to follow the good reviews and then decide whether the product is worth buying or move on to the next items. Nowadays, the drastic situations unfold as there are tensions between citizens, companies, and/or the economies; there are many unwanted reviews that are being posted online. These reviews are to encourage buyers to buy products that may be deemed to be junk, or to discourage buyers to buy products because of rising competitors.

At Gide, the startup company is trying to allow online buyers to be the judges. Gide provides rigorous review analysis and determine whether the reviews are deemed to be fraudulent or honest. In the end, the company provides detailed information of the reviews for a particular product and determine a user score for each author; the fraudulent reviews are to filtered out and only the honest reviews are to be reported to the online buyers to judge whether the product is worth buying.

2.0 The Purpose of A Manual in Software Development

When conducting and coding software development applications, the purpose of a manual (or so-called technical writing documentation) is to allow the next developers or potential customers to understand the code and the program clearly. Therefore, the persons can either use it, enhance it, or understand its use to real-world applications. Imagine a person meeting someone from a different ethnicity and background; the person may need a translator to speak to the other person. Similarly to programming, a program code is like another language that needs proper documentation and almost low-level of understanding for users to efficiently use such program.

2.1 Manual Part 1: Web Scraping

The purpose of having a web scraping manual component is to explain how to extract important information from websites that contain products and reviews. The idea is very easy however, to scrape information from websites require a lot of understanding the sites' dynamical web elements and contents in order to scrape without any issues. For instance, given the fact that a programmer will need to scrape certain information such as the title, the body text, the product URL, the web content elements will need to be extracted so the program will understand what is the source code on the website and call its contents. Therefore, the most important concepts of web scraping is to allow the developer or customer to understand what web elements to call, what tools to use, what APIs to call, and most importantly,

how to effectively run source codes without the site blocking. Obviously, there are areas that the manual should address; for example, improvements in the web scraping such as the optimization and efficiency. It is terrible if sites want to block from programs scraping their web contents and information. Furthermore, scraping website data takes a decent amount of time; to optimize the scraping efficiency will greatly enhance the runtime of retrieving web information quickly in order to get started with the analysis (manual part 2)

2.2 Manual Part 2: Review Analysis

The web scraping component is one area of work that is not the most critical component of the job; mainly Gide can use several APIs (possibly not made yet for the contents Gide wants in web scraping). This is because getting the data is not the most complicated, but analyzing product reviews is the key to determine if the reviews are valid or suspicious (also known as fake reviews). Analyzing suspicious reviews require a lot of computational/mathematical processes and logistics in order to determine certain reviews are flagged as suspicious or unreal. By applying certain statistical distributions and natural language processing, they are easily to determine the quality, the number of words, and the validity of the reviews written by specific authors. One good characteristic of a review analysis manual is the ability to allow developers or customers understand what rules and algorithms they are analyzing in the reviews not using technical language, but using basic and plain language. The other good characteristic of the manual is describing each rule/feature separately instead of jumbo them together, making things confusing.

2.3 Manual Part 3: Machine Learning and Report Output

Once the analysis component is finished, the next step is very straightforward: train the dataset of the analysis into a machine learning model for future implementations (i.e. future web scraping calls and analysis). It will categorize reviews as trolls, bots, or normal reviews; then, it will generate the trained dataset as a specific report output for test dataset from other website calls (preferably Amazon for now as this is what Gide has been using the most so far). As for the machine learning model algorithm itself, it is implemented by hand; the program itself is not called through an API. However, it is sophisticated enough to do the job as the algorithm itself is used by many statisticians and machine learning developers. Once the dataset is tested for report output, it will become a specific file format for investors, large corporations, business customers to seek and visualize its contents for their own purposes. Right now, Gide only reaches the first stage of development; there are so many other things that need to be implemented. Possible improvements will involve suggesting alternative approaches to machine learning or alternative visualization for the report output for businesses and investors to easily look at.

3.0 Gide User Manual - Initial Stage

Gide Product Analysis: This is a small-beta platform where the user will be entering a website URL that contains product reviews. As of now, it is currently only support Amazon sites. There will be 2 outputs for now (there will be more outputs in the future as we utilize Machine Learning algorithm). Important Note: This

manual contains similarities to the comments in the source code and describing the functionalities of the source code. It will also include the inputs and outputs of each functions. Make sure you are on Python 2.7 (technically the syntaxes formatted to work on Python 3.6 and above, but just on the safe side). Behind the scenes of the program, it is based on micro-service architecture, a software development technique to divide program tasks into different categories. Not only does it allow organized and readable code, but it also allows separate package calls without interfering the other program packages.

3.1 To Get Started

What the user will need for testing and usages:

1. Eclipse (Recommended): **Eclipse IDE for Eclipse Committers**
   https://www.eclipse.org/downloads/
2. PyDev on Eclipse: Navigate to **Help**. Click on **Eclipse Marketplace**. Type in **PyDev** and select **Install**
3. EGit on Eclipse: Navigate to **Help**. Click on **Eclipse Marketplace**. Type in **EGit** and select **Install**
4. Getting Terminal ready
5. Postgres Installation (server/port calling): https://postgresapp.com
6. PSequel Installation (recommended for better viewing the database):
   http://www.psequel.com

3.2 Installing Libraries and API Packages

We have a selection of libraries that need to be installed beforehand. Install them in the order of occurrences below:

1. *sudo easy_install pip*

2. *sudo -H pip install psycopg2*

3. *sudo -H pip install uuid*

4. *sudo -H pip install -U python-dateutil*

5. *sudo -H pip install -U selenium*

6. *sudo -H pip install -U DateTime*

7. *sudo -H pip install -U regex (you may not need to install this as Python contains a standard package called: re)*

8. *sudo -H pip install -U nltk*

9. *sudo -H pip install -U pandas*

10. *sudo -H pip install -U numpy*

Installing Numpy may cause issues as it overlaps with pandas.

If the error message contains numpy already existed in pandas with the latest version, you can ignore the installation process.

Otherwise, run: sudo -H uninstall numpy Then, run: sudo -H install -U numpy

11. *sudo -H pip install scipy*

12. *sudo -H pip install -U scikit-learn*

13. *sudo -H pip install matplotlib*

14. *sudo -H pip install statistics*

15. *sudo -H pip install –upgrade "watson-developer-cloud>=2.4.1"*

When installing Selenium, it is recommended to install the WebDriver for Firefox (Firefox WebdDriver). The file name is called: geckodriver-v0.xx.0-macos (https://github.com/mozilla/geckodriver/releases) where "xx" denotes version of the geckodriver. Once the download has finished, please move the executable file to folder: */usr/local/bin* (otherwise, the program won't run!)

4.0 Gide User Manual - Part 1

Part 1 of Gide user's menu deals with web scraping. The web scraping component contains two major packages: *SeleniumAnalysis* and *Extraction/Database*

4.1 Scraping Website Reviews

Scraping website reviews use the first two packages: *SeleniumAnalysis* and *Extraction/Database*. Despite *Database* is a separate package from *Extraction*, they collaborate together to gather information and store them onto the database server

using PostGresSQL protocol. The package *SeleniumAnalysis* contains five Python development modules: **automatedExt.py**, **SeleniumReviewScraper.py**, **SeleniumUserProfileScraper.py**, **reviewConfig.py**, and **webConfig.py**. The packages starts off from the module called: **automatedExt.py**; it inputs a string website URL input, and outputs the product name and the product website URL. Then, the module calls **SeleniumReviewScraper.py** module class that starts analyzing the site's web elements and extract the following information: Product name, product URL, average rating, review title, review text, date written, rating, and verified purchaser. From there, it will call another class call **SeleniumUserProfileScraper.py** that scrapes the user profiles of each review such as: Author name, author URL, rank of the reviewer, number of reviews, review title, review text, date written, and rating. As of now, the site can take only take Amazon websites.

product	product_url	author_id	author	words_found	review_text
Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	18018623…	Eldon Cooper	['quiet']	['It is exactly what I expected A…
Anker 3.5mm Nylon Braided Au…	https://www.amazon.ca/Anker-…	27356544…	makaveli	['fast', 'strong']	['Strong wire definitly buying it a…
Gentle Muzzle Guard for Dogs…	https://www.amazon.ca/Gentle-…	37599018…	Allrounder	['plastic', 'last', 'build', 'sturdy']	['Excellent Love the soft feel an…
Lorann Oils Dram 10 Pack FF#…	https://www.amazon.ca/Lorann-…	26021664…	Amazon Customer 2018-11-16	[]	[]
Cargo Net for Trailer Truck Pick…	https://www.amazon.ca/Cargo-…	25124452…	Amazon Customer 2018-11-27	[]	[]
NUOMI TV Remote Control Or…	https://www.amazon.ca/Remote…	20067547…	Matthew Wickware	['tall']	['Im using it Thats about all I ca…
Amazon Essentials Boys Boys'…	https://www.amazon.ca/Amazo…	70655102…	Jennifer	[]	[]
Vassarette Women's Comfortab…	https://www.amazon.ca/Vassar…	20422368…	Saima Tazreen	['up', 'down', 'high']	['Though this is not something I…
Wahl Canada 3145 Elite Pro Hi…	https://www.amazon.ca/Wahl-C…	22904602…	SIBY PAUL	['quality']	['Excellent quality']
Preethi Eco Plus Mixer Grinder	https://www.amazon.ca/Preethi-…	22904602…	SIBY PAUL	[]	[]
Paw Patrol, Lights and Sounds…	https://www.amazon.ca/Patrol-L…	26817696…	Amazon Customer 2016-12-27	[]	[]
adidas Women's Tastigo 15 So…	https://www.amazon.ca/adidas-…	98829279…	Amazon Customer 2018-10-23	['weight', 'light']	['Light weight thin material but n…
Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	26767465…	Amazon Customer 2018-09-14	['price']	['Its easy to assemble and good…
Earthly Body Marrakesh X Leav…	https://www.amazon.ca/Earthly-…	98829279…	Amazon Customer 2017-04-10	['smell']	['I got a product from this line in…
KAYIZU Men's Sport Performan…	https://www.amazon.ca/KAYIZ…	82141803…	Mike Philpott	['look', 'last', 'quality']	['Look great Hopefully they last…
Earth Mama Organic Nipple Bu…	https://www.amazon.ca/Earth-…	13984898…	Nida F.	[]	[]
ZYLISS 3 Piece Paring Knife S…	https://www.amazon.ca/ZYLISS…	16784408…	scott morrell	['top', 'expensive']	['We stopped using our expensi…
R.E.D. (Special Edition) (Biling…	https://www.amazon.ca/D-Spec…	25575817…	Rizzle Dizzle	[]	[]
iRULU X1S 7" Google Android…	https://www.amazon.ca/Google…	70655102…	Jennifer	[]	[]
ADHERETOFLY 50 Pcs 8 Hole…	https://www.amazon.ca/ADHER…	26021664…	Amazon Customer 2018-11-16	[]	[]
Ogilvy on Advertising	https://www.amazon.ca/Ogilvy-…	13983443…	Kindle Customer	[]	[]
DakPets Deshedding and Light…	https://www.amazon.ca/DakPet…	27356544…	makaveli	['plastic', 'last', 'part']	['Used this on my dog Does as…
Ohuhu 12-Slot Leather Watch…	https://www.amazon.ca/Ohuhu-…	17611389…	sylvie Rodeghiero	[]	[]
Majestic Earth Plant Derived Mi…	https://www.amazon.ca/Majesti…	38356748…	neighbour2thenorth	['up']	['This company is in my view pr…
Spigen Kuel A210 Metal Plates…	https://www.amazon.ca/Spigen-…	16784408…	scott morrell	['down']	['I am writing this review as i dri…

Figure 1: Words Labelling Table

4.2 Pushing to Database and Extraction

Once the above three modules finish scraping, initially the idea outputs were to be put on three .csv-formatted tables: *review_table.csv*, *product_table.csv*, *author_table.csv*. Each table contains different information of the product reviews scraped: the *review_table* contains review information such as the rating, the date reviewed, the review title and the review text, and lastly the verified purchase indicator; the *product_table* contains product information such as the product name, product URL, and the average rating; the *author_table* contains reviewer information such as author name, author URL, author rank, and the number of reviews. However, once data gets large in the future, it is better to put the tables onto the database through PostGresSQL protocol as files take up huge amounts of space in the storage.

To ensure the database is pushed correctly, the *SeleniumAnalysis* package will call

keywords	avgkeyword_confidence_score	review_sentence
scratch	[0.0, 2.75]	[["it works fine and does what it'…
cheap	[0.551080777777778, 2.14861111…	[['Prefer if it were cheaper.', 255…
cold	[0, 0]	[]
speed	[3.044288499999999, 2.29166666…	[["Bought two of these and the…
compact	[0.8653660000000001, 0.7183333…	[['Also, the packaging was very…
slow	[0.6439391999999997, 1.2422916…	[['Unfortunately, the rubber stick…
knob	[0, 0]	[]
swivel	[0, 0]	[]
returned	[0, 0]	[]
breaks	[0, 0]	[]
safe	[0.760111, 0.5583333333333333]	[['Seems safe around water(not…
front	[0.05531200000000001, 0.251851…	[['Plenty of space in the front po…
part	[2.083920187499999, 1.29361111…	[['Great for apartments.', 10788…
condition	[-4.4720005, 3.1625000000000005]	[['Well, not fast enough.. blades…
down	[-0.669501, 0.7346450617283952]	[['Unfortunately, car broke down…
cable	[0.3090155, 0.16875]	[['It comes with a rechargeable…
settings	[0.825281, 0.0]	[['Durable product, has battery t…
die	[0.641178, 0.45]	[['The whole idea of heated hoo…
fragile	[0, 0]	[]
DOD	[0, 0]	[]
small	[0.19334268292682916, 0.763159…	[['Way too small.', 1078894864…
long lasting	[0, 0]	[]
side	[0.683941052631579, 0.88411698…	[['The I nside foam is accessibl…
heavy	[0.40377471428571426, 0.591722…	[['Nice and heavy', 8214180327…
harm	[1.8675365000000002, 0.0]	[['3M tape worked like a charm.'…

Figure 2: Sentences Labelling Table

a method to push reviews from *Database* package: creating tables for the reviews information, setting up table headers and assigning them to specific types of parameters (i.e. Numeric type, Text type, etc.). Once the database tables are set and linked by primary keys and foreign keys, the extraction process begins. Other database tables include *wordslabelling* (Figure 1) and *sentencelabelling* (Figure 2) where the program will label the keywords and key sentence phrases in the reviews. Such tables look like this:

author_id	product_id	author_url	author	rank	num_of_reviews	date_scraped	user_score
29630451…	13935612…	https://www.amazon.ca/gp/profi…	s	445338	2	2018-12-18 11:59:37.535031	1000.0
82141803…	24548724…	https://www.amazon.ca/gp/profi…	M	184677	21	2018-12-18 11:52:38.55068	1000.0
12461035…	24548724…	https://www.amazon.ca/gp/profi…	m	1481604	1	2018-12-18 11:51:59.577262	1000.0
13984898…	95646868…	https://www.amazon.ca/gp/profi…	N	664868	3	2018-12-18 11:56:12.145953	1000.0
16784408…	13210336…	https://www.amazon.ca/gp/profi…	s	48034	56	2018-12-18 11:51:05.684179	1000.0
17611389…	11925584…	https://www.amazon.ca/gp/profi…	s	280968	20	2018-12-18 11:55:14.525557	1000.0
20067547…	11925584…	https://www.amazon.ca/gp/profi…	M	97841	21	2018-12-18 11:54:04.385207	1000.0
21992062…	95646868…	https://www.amazon.ca/gp/profi…	A	674450	3	2018-12-18 11:55:52.255201	1000.0
24507998…	95809880…	https://www.amazon.ca/gp/profi…	Ja	847588	3	2018-12-18 11:58:50.127557	1000.0
26052727…	23141241…	https://www.amazon.ca/gp/profi…	Je	384529	8	2018-12-18 12:01:10.02181	1000.0
27066033…	95809880…	https://www.amazon.ca/gp/profi…	D	732664	3	2018-12-18 11:58:32.041522	1000.0
32801819…	13935612…	https://www.amazon.ca/gp/profi…	V	1004599	2	2018-12-18 12:00:01.741361	1000.0
37599018…	13210336…	https://www.amazon.ca/gp/profi…	A	29235	12	2018-12-18 11:49:11.539788	1000.0
38356748…	24548724…	https://www.amazon.ca/gp/profi…	n	24223	25	2018-12-18 11:53:44.189729	1000.0
41804070…	13935612…	https://www.amazon.ca/gp/profi…	K	24168	0	2018-12-18 12:00:16.249736	1000.0
44193859…	24548724…	https://www.amazon.ca/gp/profi…	C	1498951	1	2018-12-18 11:53:25.628602	1000.0
44457885…	24548724…	https://www.amazon.ca/gp/profi…	A	153461	22	2018-12-18 11:53:03.813185	1000.0
60934998…	13210336…	https://www.amazon.ca/gp/profi…	R	319547	5	2018-12-18 11:49:52.797818	1000.0
70463209…	13210336…	https://www.amazon.ca/gp/profi…	A	72777	57	2018-12-18 11:50:01.312063	1000.0
70655102…	13210336…	https://www.amazon.ca/gp/profi…	Je	4540616	192	2018-12-18 11:50:18.655936	1000.0
78174033…	24548724…	https://www.amazon.ca/gp/profi…	C	1991580	9	2018-12-18 11:53:34.290145	1000.0
79354893…	95809880…	https://www.amazon.ca/gp/profi…	Te	506859	5	2018-12-18 11:58:09.095128	1000.0
80154574…	23141241…	https://www.amazon.ca/gp/profi…	M	48572	43	2018-12-18 12:00:55.566992	1000.0
80378263…	95646868…	https://www.amazon.ca/gp/profi…	m	320064	11	2018-12-18 11:56:39.043717	1000.0
82973586…	95809880…	https://www.amazon.ca/gp/profi…	D	1412284	1	2018-12-18 11:57:40.268536	1000.0

Figure 3: Author Table

4.3 Database Tables Viewing

The tables can be displayed easily using a simple application without calling the table through Terminal (Command) window. This application is called *PSequel*; it allows the user to setup a new connection and connect to the database to view database. The inputs are simple: the connection requires the host name, the user name, and the database where the tables are stored. Once the user is inside the database, there will be three major tables the product reviews are stored in: *author_table* (Figure 3), *product_table* (Figure 4), and *review_table* (Figure 5). All three of them will contain exactly the same format as the .csv-formatted table outputs.

The outputted tables are in the following layouts; these tables can be viewed by connecting to the database:

product_id	author_id	product	product_url	average_rating	max_rating	date_scraped
25292568…	98829279…	FIRM ABS Women's Yoga Capr…	https://www.amazon.ca/FIRM-A…	4.3000002	5	2018-12-18 11:52:47.93887
96385695…	13504966…	AmazonBasics 3-Outlet Surge…	https://www.amazon.ca/Amazo…	4.5	5	2018-12-18 11:58:41.156672
12191537…	27356544…	Bluedio T2s Bluetooth Headph…	https://www.amazon.ca/Bluedio…	4.1999998	5	2018-12-18 11:52:08.033144
39663518…	98829279…	BATTOO Anchor Theme Love…	https://www.amazon.ca/BATTO…	5.0	5	2018-12-18 11:52:47.93887
56838056…	25575817…	Castnoo 28W Full Spectrum Le…	https://www.amazon.ca/Castno…	3.7	5	2018-12-18 11:51:34.48353
72157541…	70463209…	ArtNaturals Aloe Vera Gel - (12…	https://www.amazon.ca/ArtNatu…	4.0999999	5	2018-12-18 11:50:01.312063
11925584…	17611389…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:55:14.452403
11925584…	20067547…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:04.312956
11925584…	89127004…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:55:25.662879
11925584…	90974852…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:55:33.175388
11925584…	15517061…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:55:06.380577
11925584…	16852966…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:38.353831
11925584…	22688431…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:47.29264
11925584…	23122787…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:28.942735
11925584…	26021664…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:17.907818
11925584…	26817696…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:54.073753
11933250…	89127004…	Carex Wheelchair, 1 Count	https://www.amazon.ca/Carex-…	3.2	5	2018-12-18 11:55:25.726069
15907906…	23122787…	SADES 810 PC PS4 New Xbox…	https://www.amazon.ca/Headse…	3.9000001	5	2018-12-18 11:54:28.999348
16184685…	98829279…	Women's Period Menstrual San…	https://www.amazon.ca/Menstr…	2.5	5	2018-12-18 11:52:47.93887
17460390…	25124452…	STANLEY 33-900 FatMax Extr…	https://www.amazon.ca/STANL…	5.0	5	2018-12-18 12:00:25.591612
17664536…	16784408…	GEEKHOM Tongs 9+12 Inch Ki…	https://www.amazon.ca/GEEKH…	4.6999998	5	2018-12-18 11:51:05.684179
18629787…	60934998…	iHome iAVS1B Bedside Stereo…	https://www.amazon.ca/iHome-i…	3.5999999	5	2018-12-18 11:49:52.797818
21665859…	26817696…	Comfort Zone CZST161BTE P…	https://www.amazon.ca/Comfor…	4.0999999	5	2018-12-18 11:54:54.119381
21795741…	82141803…	Yogavni Yogavni-Chequered-T…	https://www.amazon.ca/Yogavn…	4.0	5	2018-12-18 11:52:38.55068
24047968…	13983443…	JETech Screen Protector for Sa…	https://www.amazon.ca/JETech…	4.0999999	5	2018-12-18 11:52:24.402332

Figure 4: Product Table

review_id	rating	verified_purchase	max_rating	review_title	review_text	date	author_id	date_scraped
100499295664353	5	Yes	5	Five Stars	Good price and handy little gad…	2016-10-14	80154574…	2018-12-18 12:00:55.566992
100590372596657	5	Yes	5	Works great i have them place…	Works great i have them place…	2018-01-30	25124452…	2018-12-18 12:00:25.591612
100596463919065	3	Yes	5	Good	Decent mask. Goes on nice an…	2017-03-23	70463209…	2018-12-18 11:50:01.312063
100621750798252	5	Yes	5	This book is so nice and farty.	My face is full of warts and wrin…	2018-09-15	37599018…	2018-12-18 11:49:11.539788
100671918327952	5	Yes	5	Five Stars	love this book	2016-11-24	70655102…	2018-12-18 11:50:18.655936
100687575901719	4	Yes	5	Four Stars	My wife loves this no more mes…	2018-05-30	25124452…	2018-12-18 12:00:25.591612
100998969266892	4	Yes	5	Four Stars	does the trick	2018-11-02	70655102…	2018-12-18 11:50:18.655936
101000554243590	5	Yes	5	Fast shipping	These work great the kids wear…	2018-11-27	25124452…	2018-12-18 12:00:25.591612
101106515410917	5	Yes	5	Five Stars	Excellent charger. Has helped…	2017-06-09	26817696…	2018-12-18 11:54:54.119381
101325526380969	3	Yes	5	Le produit est bien. Qualité mo…	Le produit est bien. Qualité mo…	2017-03-31	78174033…	2018-12-18 11:53:34.290145
101608492988388	1	Yes	5	did not work	It was cute and my son loved it…	2018-11-02	70655102…	2018-12-18 11:50:18.655936
101767955305368	1	Yes	5	Not worth it	Had it for only one month and o…	2018-09-30	20094362…	2018-12-18 11:56:55.301185
101800447493262	4	Yes	5	Four Stars	Nice set of stories.	2017-02-10	70463209…	2018-12-18 11:50:01.312063
102012933812086	5	Yes	5	Five Stars	fits and works great	2018-11-02	70655102…	2018-12-18 11:50:18.655936
102082533005827	5	Yes	5	Still on my phone	Amazing product Does what it s…	2018-10-01	27356544…	2018-12-18 11:52:08.033144
102246730861257	4	Yes	5	... my brother who has been se…	Bought this for my brother who…	2018-09-07	20422368…	2018-12-18 11:51:21.960912
10235416613492	5	Yes	5	Good over all shoe	Could be a little more grippy bu…	2018-11-29	44457885…	2018-12-18 11:53:03.813185
102732124521786	4	Yes	5	good bag cheap price	Good product. Does not fit insid…	2018-06-01	25575817…	2018-12-18 11:51:34.48353
103188036670986	5	Yes	5	Good and fast service	Great framing hammer. Wouldn…	2018-11-27	25124452…	2018-12-18 12:00:25.591612
103526191718878	5	Yes	5	Five Stars	Love this mop and bucket!	2017-04-10	98829279…	2018-12-18 11:52:47.93887
103997782235041	5	Yes	5	works for me	Just like what grandpa used. T…	2017-09-04	25575817…	2018-12-18 11:51:34.48353
10414854101968	5	Yes	5	Great value!	Maybe the best standing fan yo…	2018-11-09	89127004…	2018-12-18 11:55:25.662879
104425742600986	3	Yes	5	Does the job	I needed something dark to filte…	2018-09-13	20422368…	2018-12-18 11:51:21.960912
104708421523018	5	Yes	5	Been using for a month now !	Have been using it for a month;…	2018-12-07	13504966…	2018-12-18 11:58:41.156672
104915942118159	3	Yes	5	Meh	Worked great at first. After a wh…	2017-12-27	70463209…	2018-12-18 11:50:01.312063

Figure 5: Review Table

5.0 Gide User Manual - Part 2

Part 2 deals with product analysis. Web scraping component contains three major packages: *SeleniumAnalysis*, *Extraction/Database*, and *FilteringAnalysis*.

5.1 Analyzing and Filtering Reviews

This phase is where all the reviews are being calculated by the rules/features given, calculate a specific score for each author, and categorize the reviews as trolls, bots, or normal. To analyze and to filter reviews are the next important steps in understanding reviews. Getting the data is the easy component, but the key goal of the data is to have a good amount of information to allow the user to understand what is happening with the reviews. Each rule/feature will be backboned by specific algorithms; not getting to the details of the algorithms for each rule, but it will be provided with a light overview of what it calculates.

Analyzing reviews will take consideration of several rules/features; the key rules and features are as follows:

1. Word Count Score; **word_count.py**

2. Quality Word Usage Score; **quality_word_usage.py**

3. Sentiment Review Score; **review_sentiment.py**

4. Word Repetition Score; **word_repetition.py**

5. Low Quality Word Usage Score; **misc_analysis.py**

6. Verified Purchase Score; **misc_analysis.py**.

These rules/features will take huge considerations of determining whether the author are writing good (normal) or bad (trolls, bots) reviews, and not the positive or the negative reviews). Furthermore, scores are calculated based on good (normal) or bad (trolls, bots) reviews. As for now, the score is currently set at 1000 (neutral). If the author's score is below 1000, his or her reviews are somewhat alright, suspicious, or terrible reviews. As for author's score is above 1000, his or her reviews are somewhat good, excellent, and awesome reviews that are deemed to be useful. However, the score parameter is subject to change as it gets confusing since the value 1000 is not a clear value or parameter that determines the reviews' validity, value, and worth; a user cannot tell whether 1100 is better than 900 or vice versa.

5.1.1 Word Count Score

The algorithm that is used here is so-called the *Zipf's Word Frequency Distribution*. It outputs a log-log distribution graph where it analyzes the frequency of the words versus the inversely proportional to the rank of the words. The end result for this rule/feature is to analyze how many words a review contains and how many unnecessary words a review puts. Then, it categorizes if the one review is using too few or too many words compared to all other reviews and deduct the author's score who wrote that particular review. Using too few words, the reviews are useless and meaningless; using too many words, the re-

views will have a lot of pointless statements. However, this is just to assume that the authors may be using a lot of unnecessary words in their reviews; in the end, the outcome is it doesn't matter if the authors are writing a short review or a long review but the contents have to be useful and meaningful.

5.1.2 Quality Word Usage Score

The algorithm that is used in this rule/feature is called *Cramer's V Distribution*. It determines the relationship between the quality words use and the reviews. For example, suppose a person buys a television; in the review, he or she will probably use some quality words that are features of the television such as dimension, pixels, LED panels, 4K, and many more other feature words. Therefore, they are to be stored in the a corpus-text file where the algorithm will search through all of them and determine if there is a relationship between the words and the reviews. The main challenge of this algorithm is to see if the word count rule can help and effect the outcome of the quality word usage score; by integrating Zipf's Word Frequency Distribution into Quality Word Usage Score and determine whether short or long reviews can have the following scenarios:

1. Writing meaningless or meaningful short reviews with few or a lot of quality words

2. Writing meaningless or meaningful long reviews with few or a lot of quality words

This rule/feature is probably one of the most complicated rule to set up and to implement.

5.1.3 Sentiment Review Score

There is no algorithm in this rule/feature however, the rule itself is calling an IBM Watson API that handles Natural Language Understanding. The API outputs the sentiment score and polarity (i.e. positive or negative) of the sentence in a review. Although the API is analyzing all the sentences of the reviews, obviously this is one of the rules where it takes a certain amount of time to run. First of all, the program uses a special API key followed by the API method to analyze sentiment for each sentence of a review. Then, analyzing one review at a time, it will call a sentence tokenizer method where it breaks the review into sentences. The list of sentences inputs into the Watson NLP API to start analyzing all of them individually and outputs the polarity and sentiment score for each sentence. Once all the reviews have been analyzed, all the sentiment scores will run through a regression model that determines if certain authors are talking mostly positive reviews, mostly negative reviews, or a mixed of both. The ultimate goal of this rule is to catch and to visualize if any authors are just talking positive reviews or negative reviews to either fool customers into buying the products, or to screw the seller, falsify and lie about the product to the customers.

5.1.4 Word Repetition Score

The rule/feature here is to determine if there are any reviews contain words that are repeated in their other reviews written by the same author. So far, there is no algorithm; it is a simple calculation-based rule that neglects any non-useful words such as ”a”, ”the”, ”I” - words that portray no meaning, and analyze meaningful words that are used repeatedly throughout the same author’s reviews. Scores are deducted based on if the reviews are repeated or if there are words that are repeated consistently. Repeated reviews or words in the reviews are caught to be either trolls or bots trying to lie the quality and the product information to the customers.

5.1.5 Low Quality Word Usage Score

The rule/feature here is to determine if there are any slang words or profanities used in the reviews. A simple corpus-based text that stores all the possible profanities and slang words, and analyze all the reviews to determine if any authors are using them in their reviews. Scores are deducted based on how many slang words and/or profanities used in author reviews.

5.1.6 Verified Purchase Score

The last rule/feature is extremely straightforward: to determine verified purchasers from the reviews. If a review has a verified purchaser badge, the author

of the review is considered to be the customer who actually bought the product. For reviews that are verified, customers tend to trust it more than non-verified reviews. Therefore, scores are deducted based on how many non-verified purchase reviews the author have put on.

6.0 Gide User Manual - Part 3

Part 3 of Gide user's menu deals with machine learning training and report outputs for visualization. Machine learning component contains one major package: *MLAnalysis*.

6.1 Phase 5: Machine Learning Training Models

Once the analysis is completed, the dataset and the scores are being stored in a special file called *dataset_train.csv*. Here, there is a new column created called *Category*. So far, the *Category* column contains several categorizations: *troll*, *bot*, *normal*, etc. ; the categorizations are determining whether the reviewers are trolling their reviews, automating their reviews, or writing normal and honest reviews respectively - there are many other categorizations. The categorizations are to be inputted manually, but in the future, once the dataset gets very large, most of the reviews are trained and categorized so they do not need to be retrained again.

To train the dataset: once all the reviews are categorized, the program uses k-means neighbours and linear discriminant analysis machine learning models to simply cat-

egorize the newly outputted dataset from web scraping and review analysis by using the original trained review dataset *dataset_train.csv*. The two models understand each other where the categories lie based on the scores given in the trained review dataset and determine classification rules for categorizations (linear discriminant analysis). Then, it will approximately classify the newly outputted dataset into different categories (k-means neighbours).

To understand what each model does, k-means neighbours machine learning model is a clustering analysis where the dataset are clustered together based on similarities; if there are new inputted review data coming in, they will find the nearest neighbour dataset that have similarities with them; therefore, the review data are classified into different categories. The uses of k-means neighbours have similarities with linear discriminant analysis: they are both machine learning analysis in modelling the dataset. However, k-means clustering is an unsupervised training method where the categories are unknown (there is no one to tell the machine learning model what to do), whereas linear discriminant analysis is a supervised training method where the categories are manually inputted (in other words, there is someone to tell the machine learning model what to do). Despite that k-means is unsupervised and therefore less accurate, it is proposed to combine k-means with linear discriminant analysis in order to achieve semi-supervised machine learning method. This achieves efficiency as the accuracy will improve once the dataset increases. Together, linear discriminant analysis can manually label the categories and automatically produce new dataset with categorizations, and create new classification rules to support k-means to better classify reviews into categories.

6.2 Phase 6: Report Outputs

At the current state, the report outputs are not from the machine learning training models, but rather from the review analysis. This is because report outputs are required to train for future report outputs where categorization are inputted automatically by the machine learning models. Nevertheless, the report outputs are very straightforward. In a business perspective, the reports are crucial to the company's success.

The first report, *product_info.csv* displays feature keywords gathered from a unique product URL, follow by the keyword sentiment and confidence scores, the sentences that the keywords have found, the sentences sentiment and confidence scores, and the author name. Most importantly, it also shows the review ratings to compare between the sentences of the reviews and the ratings of the reviews.

The second report, *sentiment_info.csv* displays the ratings and average ratings from a unique product URL, follow by the author name, the review text, the date of the review, the verified purchase identifier (is the review from a verified purchaser?), the number of high quality words (feature words for the particular product), the number of low quality words (slang words and profanities), and the number of words in the review text. Most importantly, it also shows the sentiment score of the entire review text to compare between the text, the number of high and low quality words, and the ratings.

Lastly, the third report, *author_info.csv* displays the author name, the number of reviews the author wrote, the product URL, the ratings/average ratings of that particular product, the review text/review date of that particular product, the sentiment score of the review text, the number of low quality words, the number of words used in the review, and the verified purchase identifier. Most importantly for this table, it collects all the reviews not one particular product but all the products the authors have written.

The reports are generated easily by selecting specific data information from the tables in the database. Since there are three reports to be produced, the data information will have three data-structured lists. Then, the lists will store the information onto the .csv output by writing a file to a designated output folder. Once the reports are in, developers and/or customers can use them to create visualizations to better visualize the data and manually categorize them for machine learning purposes. In the future, the reports will automatically become visualizations for displaying the results of the product reviews.

6.2.1 Score Output

Report outputs are generated from review analysis during first round of website scraping. However, in order to produce dynamically, unsupervised, or obtain accurate categorizations for future report outputs, scores are given for each author based on their review qualities. The scores are given to train new datasets (reviews) and categorize them to output new reports.

word_count_score	quality_word_score	review_sentiment_score	word_reptition_score	word_repetition_percentage	low_quality_score	verified_purchase_score	review_text
995.632959198091	789.971199358299	1006.58015727006	950.0	0.5	1000.0	2000.0	{{'''Good product'',0.0},{'''Good…
971.79409870147	177.479920852453	1755.90081329383	1006.25	0.0625	1000.0	2000.0	{{'''Great filters.'',0.0},{'''Great pr…
989.350587026306	854.556611899589	1006.58015727006	1007.69230769231	0.0769230769230769	1000.0	2000.0	{{'''it works fine and does what it…
998.370141873159	730.267002941305	1311.6997672176	1014.28571428571	0.142857142857143	1000.0	2000.0	{{'''A good buy for the money I p…
800.271981714126	11.3400221980093	1558.99349351103	1002.32558139535	0.0232558139534883	1000.0	2000.0	{{'''Not magnetic so it didnt work…
849.386864784831	191.990087698234	1253.38639273382	1016.66666666667	0.166666666666667	1000.0	2000.0	{{'''So cute'',0.0},{'''Work perfect…
930.964577881391	177.479920852453	657.573838924495	1008.33333333333	0.0833333333333333	1000.0	2000.0	{{'''Nothing fancy butworks like…
999.77165919648	730.267002941305	1461.65748608381	1011.11111111111	0.111111111111111	1000.0	2000.0	{{'''Good fan, pretty loud on the…
994.019492662974	730.267002941305	1176.5393864877	933.333333333333	0.333333333333333	1000.0	2000.0	{{'''It's good. No problems.'',0.0}…
951.493551757124	492.985078524379	1267.31887659645	1007.14285714286	0.0714285714285714	1000.0	2000.0	{{'''Not even 24 hrs of use and a…
999.925945796692	730.267002941305	1202.27834281932	950.0	0.5	1000.0	2000.0	{{'''Fan works great, as it should…
998.002510245644	789.971199358299	1006.58015727006	1000.0	1.0	1000.0	2000.0	{{'Good',0.0},{'Good',0.0},{'Goo…
982.295167759903	360.010735788783	986.238738329127	1002.22222222222	0.0222222222222222	1000.0	2000.0	{{'''I got what I paid for. The plas…
935.531637238057	129.607729883181	1083.30789211102	1005.0	0.05	941.176470588235	2000.0	{{'''Quiet and extremely lightwei…
999.975314656203	924.422312527986	1107.23817299706	933.333333333333	0.333333333333333	1000.0	2000.0	{{'''This is a great working fan.…
998.290222722016	854.556611899589	1006.58015727006	1010.0	0.1	1000.0	2000.0	{{'''Not the best time quality you…
978.438642713288	164.066398861708	2361.62586779421	1007.69230769231	0.0769230769230769	1000.0	2000.0	{{'Amazing',0.0},{'''Great produc…
998.531260969253	624.054495815589	1290.17224885626	1003.44827586207	0.0344827586206896	1000.0	2000.0	{{'''I doubt anyone could make a…
890.486112459451	10.4829695444025	376.265120656456	1001.17647058824	0.0117647058823529	1000.0	2000.0	{{'''Does the job no problem .'',0…
545.690550284218	0.00025881589244660…	306.355269883808	1001.11111111111	0.0111111111111111	1000.0	2000.0	{{'''was looking for an economy…
1000.0	455.726406331298	1385.27672235316	1003.33333333333	0.0333333333333333	1000.0	2000.0	{{'''Ce produit est parfait côté qu…
992.849068899805	624.054495815589	1421.02197902824	950.0	0.5	1000.0	2000.0	{{'''Quick and easy to put togeth…
976.204539640886	31.4991223057928	412.916878212405	1001.6393442623	0.0163934426229508	1000.0	2000.0	{{'''Awesome little flashlights'',0.…
997.977874210568	389.444013747649	1774.34882643103	1020.0	0.2	1000.0	2000.0	{{'''Works perfect'',0.0},{'''Works…
999.950629921772	854.556611899589	1107.23817299706	1020.0	0.2	1000.0	2000.0	{{'''Awesome fan though the par…

Figure 6: Score Output for Categorization

Product Info	Product Name: || Date Scraped: [datetime.datetime(2018, 12, 18, 12, 0, 25, 591612)]		Product URL: https://www.amazon.ca/Comfort-Zone-CZST161BTE-Pedestal-Fan/dp/B004IPA774/ref=cm_cr_arp_d_product_top?ie=UTF8						
KEYWORDS	Keywords Sentiments Score	Keywords Confidence Score	Sentence	Sentiment Score	Confidence Score	Author	Rating	Average Rating	Category
SCRATCH	0	2.75	It works fine and does what it's supposed to do, but has scuffs and scratch marks on it.	0	0.5	meriame	2	4.1	NULL
CHEAP	0.551080777777778	2.148611111111113	The last one we bought (cheap junk from Canadian Tire) lasted 10 years and it was literally only shut off to clean it.	-0.962563	0.616666666667	Ron Seigel	1	4.1	NULL
CHEAP	0.551080777777778	2.148611111111113	Great for a cheap fan for my home gym.	0.801091	0.725	Jessica McBride	4	4.1	NULL
CHEAP	0.551080777777778	2.148611111111113	Does the trick buy feels especially cheap.	-0.730286	1	Christopher Darling	4	4.1	NULL
CHEAP	0.551080777777778	2.148611111111113	This one is maybe slightly below in terms of quality of materials our cheapest fan.	0.804172	0.166666666667	Christopher Darling	4	4.1	NULL
CHEAP	0.551080777777778	2.148611111111113	It's a cheap fan that works fine.	0.669177	0.6	Christopher Darling	4	4.1	NULL
SPEED	3.04428849999999	2.2916666666666665	Bought two of these and the motor on one of them doesn't run smoothly - it still sends breeze but it does faster and slower motion constantly regardless of the speed choice.	0.553507	0.416666666667	Allrounder	2	4.1	NULL
SLOW	0.643939199999997	1.2422916666666666	Bought two of these and the motor on one of them doesn't run smoothly - it still sends breeze but it does faster and slower motion constantly regardless of the speed choice.	0.553507	0.416666666667	Allrounder	2	4.1	NULL
PART	2.0839201874999999	1.2936111111111108	Easy to put together, works fine for cooling my small apartment	0.942883	0.577777777778	Dan McKelvey	4	4.1	NULL
PART	2.0839201874999999	1.2936111111111108	Awesome fan though the parts are not of high quality.	0.988448	0.77	Denny Samuel Thomas	5	4.1	NULL
PART	2.0839201874999999	1.2936111111111108	I worked in the department that sold fans at Sears for 4 years.	0.628221	0	Christopher Darling	4	4.1	NULL
CONDITION	-4.4720005	3.162500000000005	Well, not fast enough.. blades are just moving fast but not providing any good air conditioning.	-0.813091	0.575	Amazon Customer 2018-10-22	1	4.1	NULL
DOWN	-0.669501	0.7346450617283952	The legs fold up when you left the base to move it and when you go to set it back down if youre not careful it'll fall over	-0.60512	0.457777777778	scott morrell	3	4.1	NULL
SMALL	0.19334268292682916	0.7631597687542813	Easy to put together, works fine for cooling my small apartment	0.942883	0.577777777778	Dan McKelvey	4	4.1	NULL
SMALL	0.19334268292682916	0.7631597687542813	A little flimsy looking looking, I wouldnt have it around small children	-0.882675	0.45	Jessica McBride	4	4.1	NULL
SIDE	0.6839410526351579	0.8841168843801424	Decent quality considering the price	0.813103	0.666666666667	M.Jean Sommerdyk	4	4.1	NULL
HARM	1.8675365000000002	0	Nothing fancy butworks like a charm.	0.522418	0	Matthew Wickware	5	4.1	NULL
WEIGHT	2.3431513333333327	1.6849999999999998	Quiet and extremely lightweight	0.94625	0.666666666667	neighbour2thenorth	5	4.1	NULL
BACK	-0.5023726315789476	0.4854532163742689	The legs fold up when you left the base to move it and when you go to set it back down if youre not careful it'll fall over	-0.60512	0.457777777778	scott morrell	3	4.1	NULL
DESIGN	0.2924628333333334	0.3604761904761905	Does what is designed.	0	0	Shopaholic	4	4.1	NULL
NOISE	0.5564005999999996	3.0777777777777775	Luckily for us it's really just meant for white noise in our bedroom.	0.769989	0.344444444444	Ron Seigel	1	4.1	NULL
NOISE	0.5564005999999996	3.0777777777777775	Make a weird noise when on rotation	-0.576895	1	makavelli	1	4.1	NULL
NOISY	2.8273176666666667	3.0666666666666664	- Very easy to assemble - Not to noisy - Good quality for the price	0.855677	0.8	Christian	5	4.1	NULL
LAST	-0.36019335999999999	0.9146891774891177	It will do the job I need, but the plastic cracked easily.	0.01722	0.833333333333	Amazon Customer 2018-09-30	4	4.1	NULL
LAST	-0.36019335999999999	0.9146891774891177	The plastic is so flimsy, it is a dollar store quality.	-0.611948	0	Allrounder	2	4.1	NULL
LAST	-0.36019335999999999	0.9146891774891177	The last one we bought (cheap junk from Canadian Tire) lasted 10 years and it was literally only shut off to clean it.	-0.962563	0.616666666667	Ron Seigel	1	4.1	NULL
LAST	-0.36019335999999999	0.9146891774891177	No way this one lasts a year.	0	0	Ron Seigel	1	4.1	NULL
QUIET	2.2643812857142858	1.4785714285714282	Easy to assemble, works well, quiet.	0.916166	0.583333333333	Kindle Customer	5	4.1	NULL
QUIET	2.2643812857142858	1.4785714285714282	A nice quiet fan to circulate the air.	0.985583	0.666666666667	Elton Cooper	5	4.1	NULL

Figure 7: Product Info Report

6.2.2 First Report

For visualization purposes, although incomplete, the report in Figure 6 below is for one unique product and one unique URL:

Sentiment Info	Product Name: || Date Scraped: [datetime.datetime(2018, 12, 18, 12, 0, 25, 591612)]		Product URL: https://www.amazon.ca/Comfort-Zone-CZST161BTE-Pedestal-Fan/dp/B004IPA774/ref=cm_cr_arp_d_product_top?ie=UTF8						
Rating	Average Rating	Author	Review Text	Review Date	Verified Purchase	# HQ Words	Word Count	# LQ Words	Sentiment Score
5	4.1	SG	Good	2018-10-26	Yes	0	1	0	0
5	4.1	Amazon Customer	As advertised	2018-11-16	Yes	0	2	0	0
5	4.1	Client d'Amazon	Ce produit est parfait côté qualité-prix. Il est facile à assembler, pas trop bruyant. Il fait bien le travail!	2018-11-15	Yes	0	18	0	0.33216766666666664
1	4.1	Amazon Customer	Well, not fast enough.. blades are just moving fast but not providing any good air conditioning.	2018-10-22	Yes	1	16	0	-0.813091
5	4.1	Amazon Customer	It worked, it was easy to put together and very happy.	2018-09-05	Yes	0	11	0	0.991871
5	4.1	Amazon Customer	Maybe the best standing fan you can get at this price.	2018-11-09	Yes	1	11	0	0.693127
5	4.1	Sydney	Really easy to set up and runs great throughout the night	2018-11-13	Yes	1	11	0	0.977534
4	4.1	Vino	Good	2018-09-21	Yes	0	1	0	0
3	4.1	Bala	Good	2018-10-20	Yes	0	1	0	0
5	4.1	K-Mac	This is a great working fan. Exactly what I wanted. Easy to put together	2018-09-18	Yes	0	14	0	0.6287006666666667
4	4.1	Amazon Customer	Unfortunately the product arrived damaged but it still works. It will do the job I need, but the plastic cracked easily.	2018-09-30	Yes	1	21	0	0.0657045
3	4.1	Amazon Customer	Good for the money. Box had been opened and was shipped open. Works well	2018-10-31	Yes	0	14	0	0.30026533333333333
4	4.1	Shopaholic	Does what is designed. Build could be improved	2018-12-08	Yes	0	8	0	0.438515
2	4.1	Typhanie	It is extremely flimsy, poor quality	2018-11-12	Yes	1	6	0	-0.983938
5	4.1	Tausif Ahamed Bhuiyan	Very good fan	2018-12-04	Yes	0	3	0	0.987657
5	4.1	Amazon Customer	Very noisy but if that doesn't bother u it's a great fan	2018-11-27	Yes	1	12	0	-0.461953
5	4.1	Denny Samuel Thomas	Awesome fan though the parts are not of high quality. It serves the people	2018-10-17	Yes	2	14	0	0.831029
3	4.1	Client d'Amazon	Bien mais fait un bruit quand il arrive tout à gauche en oscillant	2018-12-06	Yes	0	12	0	-0.348935
4	4.1	Saima Tazreen	Urgently needed this for the scorching heat this summer. Arrived on time and was quite easy to set up.	2018-09-07	Yes	2	19	0	0.247221
4	4.1	Rene Landry	À la maison pour se rafraîchir.	2018-09-16	Yes	0	5	0	0.489337

Figure 8: Sentiment Info Report

Author Info										
Author	# Of Reviews	Product URL	Rating	Average Rating	Review Text	Review Date	Sentiment Score	Word Count	# LQ Words	Verifi
Amazon Customer	56	https://www.amazon.ca/Bluetooth-Headphones-AUKEY-Sweatproof-Smartphones/dp/B014R2MGDA?ref=pf_vv_at_pdctrvw_dp	5	3.7	Easy to use and fast shipping thanks	2016-09-29	0.985707	7	0	Yes
Kindle Customer	46	https://www.amazon.ca/%C3%A2mes-mortes-French-Nikolai-Gogol-ebook/dp/B00BU8JPE8?ref=pf_vv_at_pdctrvw_dp	5	5	Excellente traduction, lecture agréable.	2018-08-21	0	4	0	Yes
scott morrell	56	https://www.amazon.ca/Hayward-SP1420-Hydrostream-Fitting-Directional/dp/B004VTGN8E?ref=pf_vv_at_pdctrvw_dp	4	3.5	The only bummer is that the kids like to fidget with it.	2015-02-09	-0.916379	12	0	Yes
Jennifer	192	https://www.amazon.ca/Philips-HP6341-00-LadyShaver-Electric/dp/B009T0537W?ref=pf_vv_at_pdctrvw_dp	4	4	Smaller then expected but work ok	2017-09-19	0.449581	6	0	Yes
Michael Hudson	43	https://www.amazon.ca/Sawyer-Products-SP128-Filtration-System/dp/B00FA2RLX2?ref=pf_vv_at_pdctrvw_dp	5	4.7	Great shipping and product as shown. Thanks!	2016-12-21	0.3931085	7	0	Yes
Saima Tazreen	33	https://www.amazon.ca/Revlon-RVDR5202F-Essentials-1875-Watt-Ionic/dp/B01M9D1GDS?ref=pf_vv_at_pdctrvw_dp	5	4	I am so in love with this!!! I was a little unsure about this as I am very sensitive about my hairstyling. This has served me better than any other dryer I have used before. And the price is very reasonable! I strongly recommend t	2018-09-07	0.5678596	45	0	Yes
Sheila Borden	24	https://www.amazon.ca/European-Style-Newsboy-Retro-Irish/dp/B06ZZ1J75Y?ref=pf_vv_at_pdctrvw_dp	5	4.6	My Husband loves this hat. Fits as expected 24" his size He's planning on ordering another....	2018-02-25	0.4798095	16	0	Yes
Michael Hudson	43	https://www.amazon.ca/Stansport-Blend-Blanket-Olive-Green/dp/B00AD6KDVC?ref=pf_vv_at_pdctrvw_dp	5	4.5	Good shipping and heavy blanket. It does have a smell as others mentioned, but I just got this to use under my sleeping bag.	2016-10-14	0.2962825	24	0	Yes
AJ	22	https://www.amazon.ca/Bergamot-Essential-Oil-Calabrian-Therapeutic/dp/B015RU21KO?ref=pf_vv_at_pdctrvw_dp	4	4.3	Not as good as others on its own but it's great to mix to add richer scent.	2017-05-17	0.619548	17	0	Yes
Michael Hudson	43	https://www.amazon.ca/ThruNite-Ti3-Flashlight-Max120-Neutral/dp/B00EC0FY1O?ref=pf_vv_at_pdctrvw_dp	5	4.4	Awesome little flashlights	2016-10-14	0.860502	3	0	Yes
Christian	3	https://www.amazon.ca/GETIHU-Portable-External-Powerbank-Compatible/dp/B06Y4NXHB2?ref=pf_vv_at_pdctrvw_dp	5	4.6	Produit très efficace pour son prix. Il faut simplement s'assurer d'utiliser le câble qui est fourni avec le chargeur portatif, car avec d'autres câbles le temps de charge peut être est plus long.	2018-06-28	0.3930445	33	0	Yes
Jennifer	192	https://www.amazon.ca/Brighter-Child-Flash-Cards-Alphabet/dp/0769646794?ref=pf_vv_at_pdctrvw_dp	5	4.6	as expected	2018-11-02	0	2	0	Yes
neighbour2thenorth	25	https://www.amazon.ca/Majestic-Earth-Plant-Derived-Minerals/dp/B00IM5E0CE?ref=pf_vv_at_pdctrvw_dp	5	4.6	This company is, in my view, probably one of the premier suppliers of minerals. Although, next time I will order cherry flavour as the non-flavoured product is quite bitter. Dr. Joel Wallach is the creator of this product, I believe	2018-03-08	0.25237425	45	0	Yes
Joshua Kennedy	8	https://www.amazon.ca/Anjou-Aromatherapy-Peppermint-Eucalyptus-Humidifier/dp/B07D48YM4S?ref=pf_vv_at_pdctrvw_dp	5	4.3	I love how it works and all the scents!	2018-12-06	0.987488	9	0	Yes
Christopher Darling	10	https://www.amazon.ca/Samsung-Galaxy-X-Doria-Defense-Shield/dp/B06XYJNKQB?ref=pf_vv_at_pdctrvw_dp	4	4.4	Looks great! But the clear back already is very scuffed up after only one day of use. Also, the top part of the purple metal is a different colour. Much closer to pink than purple. It looks like it's been worn out or something. I ch	2018-03-29	-0.04172400000000002	49	0	Yes
makaveli	43	https://www.amazon.ca/SanDisk-Ultra-microSDXC-Adapter-SDSQUAR-064G-GN6MA/dp/B073JYVKNX?ref=pf_vv_at_pdctrvw_dp	1	4.6	Very upset and dissatisfied. Uploading my pictures to this memory card from a trip we made all erased. Iuno how or what happened but no picture got saved....	2018-10-01	-0.7232646666666667	28	0	Yes
Kindle Customer	46	https://www.amazon.ca/Gallipoli/dp/B074JV6W9R?ref=pf_vv_at_pdctrvw_dp	5	4.7	Saw the movie when it came out so many years ago. I enjoyed it even more now.	2018-08-06	0.609658	17	0	Yes
Rizzle Dizzle	60	https://www.amazon.ca/Lucy-Sous-titres-fran%C3%A7ais-Scarlett-Johansson/dp/B00M4Q7E84?ref=pf_vv_at_pdctrvw_dp	5	4.4	Good quality quick shipping	2017-09-04	0.904536	4	0	Yes
Amazon Customer	55	https://www.amazon.ca/Ekouaer-Maternity-Nightgown-Breastfeeding-Sleepwear/dp/B074RD36H6?ref=pf_vv_at_pdctrvw_dp	3	1.6	Good quality fabric but the night dress looks different than the pic. Shoulders are puffy - 1980's looking. Slots in the breasts are huge. Mine fall out and I am only small chested. Someone who has larger breast would never b	2018-01-11	-0.10361440000000004	42	0	Yes
AJ	22	https://www.amazon.ca/Comfort-Zone-CZST161BTE-Pedestal-Fan/dp/B004IPA774/ref=cm_cr_arp_d_product_top?ie=UTF8	5	4.1	Good product for the price. I buy one every year or 2	2018-11-14	0.3769135	12	0	Yes
Jennifer	192	https://www.amazon.ca/Vintage-Embroidery-Chinese-Embroidered-Walking/dp/B01M5ARQSD?ref=pf_vv_at_pdctrvw_dp	1	1	did not fit and could not get a hold of seller to fix the issue do not waist your money	2018-11-02	-0.907598	20	0	Yes
Louis-Étienne Gagnon	4	https://www.amazon.ca/Viper-Graphstrike-Fiberglass-Graphite-Composite/dp/B00VBL468Y?ref=pf_vv_at_pdctrvw_dp	5	3.8	This product is great for casual to advanced players. It's really cheap and good quality. The tip is in wood though and I'm a little bit surprised by that, thought it would be in leather. Good quality/price and as they say in the d	2018-09-15	0.69126	49	0	Yes
Amazon Customer	64	NO URL DETECTED	2	0	The 10 foot cord stopped working within a month of purchase.	2018-03-21	0	11	0	Yes
SIBY PAUL	11	https://www.amazon.ca/F-0-Amplifier-Installation-Wiring-Install/dp/B01I56R7OC?ref=pf_vv_at_pdctrvw_dp	5	5	Excellent stuff	2018-12-15	0.988381	2	0	Yes
Rizzle Dizzle	60	https://www.amazon.ca/Splitter-Adapter-Converter-Extender-100-230V/dp/B0784FWLGQ?ref=pf_vv_at_pdctrvw_dp	5	5	Well made product. Strong.	2018-02-03	0.3547025	4	0	Yes
Rene Landry	9	https://www.amazon.ca/Harry-Potter-Complete-Collection-Bilingual/dp/B005Q6TB44?ref=pf_vv_at_pdctrvw_dp	5	4.1	You should watch the whole serie! Very good!	2016-02-22	0.389786	8	0	Yes
Jennifer	192	https://www.amazon.ca/LEGO-Star-Wars-Speeder-75099/dp/B00WI0QCAM?ref=pf_vv_at_pdctrvw_dp	4	4.5	as expected	2017-09-19	0	2	0	Yes
Rizzle Dizzle	60	https://www.amazon.ca/Ronegye-Preservation-Double-sided-Channels-Commercial/dp/B075ST849G?ref=pf_vv_at_pdctrvw_dp	4	4.5	Good product. Does not fit inside my vacuum sealer.	2018-06-01	-0.389832	9	0	Yes
AJ	22	https://www.amazon.ca/Omron-Healthcare-Pressure-Monitor-Bluetooth/dp/B073HFBJ1H?ref=pf_vv_at_pdctrvw_dp	5	4.5	Great product and I trust it's accuracy	2018-11-14	0.985409	7	0	Yes

Figure 9: Author Info Report

6.2.3 Second Report

Figure 7 is for one unique product and one unique URL:

6.2.4 Third Report

Figure 8 is for all products that the authors have reviewed: