

Date: May 2, 2018

Bassam Kaaki

Fake Reviews – Project

Management 647

<b>INTRODUCTION</b>	<b>1</b>
RESEARCH SUBJECT AND OBJECTIVE	1
HOW DO COMPANIES DEAL WITH FAKE REVIEWS?	1
PROJECT KEY PERFORMANCE INDICATORS	2
<b>STEPS FOR FAKE REVIEW ANALYSIS</b>	<b>2</b>
1. GOOGLE WEB SCRAPING	2
2. DATA CLASSIFICATION	3
3. USING SPSS MODELER TO CREATE THE MODELS	3
<b>SPSS MODELER RESULTS</b>	<b>5</b>
CLASS NUGGET RESULTS	5
CLASS WITHOUT REVIEWS NUGGET RESULTS	6
REVIEWS NUGGET RESULTS	7
CLASS WITH REVIEWS NUGGET RESULTS	7
COMPARISON AND SUCCESS RATE FOR MODEL PREDICTION RESULTS	9
<b>CHALLENGING ASPECTS AND HOW TO DO BETTER THE NEXT TIME</b>	<b>9</b>
PYTHON COMMANDS	9
WEB SCRAPING	9
CLASSIFYING DATA (TRUE/FALSE)	10
SPSS SYNTAX ERRORS	10
HUMAN PREDICTION ERRORS	10
<b>CONCLUSION/RECOMMENDATIONS</b>	<b>10</b>
SUGGESTIONS FOR ONLINE RETAIL COMPANIES	11
<b>PROJECT SUCCESS FACTOR</b>	<b>12</b>
<b>APPENDIX</b>	<b>13</b>
<b>BIBLIOGRAPHY</b>	<b>21</b>

## **Introduction**

Many online retailers are having problems with fake online reviews that are filling up their product review pages with 5-star ratings and fraudulent text that is deceptive and not useful to customers searching for information upon purchasing online items. Customers usually look at product reviews to gain knowledge and compare prices, functionality, durability, benefits and satisfaction achieved from these products that were purchased previously by others.

## **Research subject and objective**

This research project will focus on creating a classification model using SPSS modeler, to determine fake reviews, which can positively help start-ups, small-businesses or individuals having their own websites from falling into the trap of non-realistic product review information. The project will focus on the CHAID algorithm that will classify the data collected from Amazon on a women's handbag; 1,000 customer comments will be included in this research project.

## **How do companies deal with fake reviews?**

Businesses such as Google, Amazon or Yelp as examples deal with fake reviews by flagging them on their systems using their own algorithms that are not 100% efficient. Unless they have a real spamming algorithm that actually understands a clear spamming pattern (behavioral) with at least 90-100% efficiency, they will continue to suffer with fake reviews swamping their websites. The flagged reviews usually performed by other customers can't effectively report the reviews either and therefore there is little one can do about it. As such, behavioral analysis of text is a very difficult task, since companies will have to have a very huge database of words that is in need to be classified first into fake and non-fake categories and perhaps include the mood of the person when such words were used at the time of writing the customer reviews. The mood of the person at the time of writing the comment is not an easy task to distinguish, unless the person is cursing, shows big letters, is placing words such as angry, ugly, horrible, mad etc.

## **Project Key Performance indicators**

The key performance indicators that will measure the success of this project will revolve around the below: (See appendix figure 1.0 – KPI's)

1. Categorization of 1,000 amazon comments with a weight of 20%
2. Effective model output with at least 80% accuracy with a weight of (35%)
3. 1% success rate in change when comparing (class without reviews) with (class with reviews weighted at (5%))
4. Python language knowledge with a weight of 20%
5. Google web scraping knowledge with a weight of 20%

## **Steps for fake review Analysis**

### **1. Google web scraping**

The first step in creating an excel data sheet is to perform, web scraping which involves the installation of a plug-in called "Web Scraper" on the Google browser. The Amazon link [https://www.amazon.com/Scarleton-Large-Drawstring-Handbag-H107804/dp/B009DKKUP8/ref=sr\\_1\\_sc\\_3?ie=UTF8&qid=1519675780&sr=8-3-spell&keywords=scaleton+bag](https://www.amazon.com/Scarleton-Large-Drawstring-Handbag-H107804/dp/B009DKKUP8/ref=sr_1_sc_3?ie=UTF8&qid=1519675780&sr=8-3-spell&keywords=scaleton+bag)

containing the product reviews for a (women's bag), is then inserted in an ID URL field in the web scraper program to form a site map. The selectors button will choose the required data for each field you want to extract the data for. For the scope of this project, the ID of the customer, the customer reviews and how helpful the review was to others were the fields that were selected. The next step is to paginate, i.e., extract all the relevant data in all the rest of the review pages. As the site map is created, the selector graph button is used to ensure that the required information is set up correctly and ready for extraction. (See appendix figure 1.1). Finally, all the selected data is extracted and presented in an excel file format ready for classification. The source code generated for the data extraction is disclosed in the appendix (Web scraping code document).

## 2. Data classification

Now that we have all the review comments for the Amazon handbag, the comments must be given a class of true or false. In order to do that, we must carefully read each comment and based on the below set of rules decide if its a fake or true comment using our human sentiment.

- 1) Very descriptive comments using height, width and texture name
- 2) "Marketing speak", "Selling tone", "Pushing to buy tone"
- 3) Generic comments such as like it, love it, adore it, nice, excellent
- 4) Grammatical errors in the comments
- 5) Comment does not describe why the reviewer likes or dislikes product
- 6) Helpful comments that enabled others to decide on its purchase
- 7) Over use of the product name, its brand and company name
- 8) Relating products to other brands
- 9) Same ID's commenting more than once
- 10) Giving immediate 5-star ratings in a comment

After classifying each comment according to the set of rules, we look at the attributes that were mostly used in these comments and check them against each comment. If an attribute is used in the comment, we place the number "1" and if its not used a number "0". The reason for performing such action is to enable SPSS text analytics to transform the unstructured review texts and turn them into quantitative data, in order for us to gain insight using sentiment analysis. (Please see appendix figure 1.2 – Amazon scraping sheet).

## 3. Using SPSS modeler to create the models

SPSS modeler is a text analytics and data mining software and is widely used to build what is called a predictive model and other statistical analysis features such as means, standard deviations, tree diagrams and probability to name a few. In order to start the analysis, it is vital to understand the nodes that were used to develop the model. The first step is to drag the excel node to the SPSS canvas which will import the amazon scraped file information with the 13 fields to be examined

(Please see appendix figure 1.3 – SPSS Excel node). The second step will require a Type node that will specify the field attributes (Reviews, Helpful comments, Short Message, Long Message, Quality, Size, Smell, Color, Love it, Like it, Problem Gift, Delivery) and characterize the data in measurement form such as either continuous (numerical values), categorical (string values), flag (fields that contain true/false or 0 and 1), nominal (data with multiple distinct values such as small/medium/large) or ordinal (requires a specific order such as normal/high/low), typeless (data not conforming to any of the above field attributes, of which has a mix of measurements) and finally target (the field that will be analyzed, in this case the class field where true/false is displayed. (Please see appendix figure 1.4 – Type Node). To check for the total data count of false and true comments in the class field predicted by my own examination, the distribution node can be used to view it. (Please see appendix figure 1.5 – Distribution\_class). The next step is to add the CHAID node known as the (Chi-squared Automatic Interaction Detection). This node will test for cross tabulations between the 13 attributes (predictors) and show which categories were most statistically significant in terms of predictor importance towards a false or true message based on the class field as a target. A CHAID tree model (nugget) will be generated to show the predictor importance and its rules. (Please see appendix figure 1.7 – CHAID class tree model). An analysis node is attached to the CHAID tree model that was generated to show us the overall accuracy of this model. A table node is attached to the CHAID tree model in order to view the prediction model and the confidence level of the prediction. Now its time to analyze the amazon reviews themselves by using the text mining node, which utilizes linguistic and frequency techniques, in order to pull key concepts from the amazon reviews and create categories associated with them (Please see appendix figure 1.8 – Text mining node). The ID field will contain the customer ID's, those who typed the reviews and the text field will contain the reviews (comments) for those particular ID's. We will need to partition the data ie., split it into separate subsets in order to test and validate the (Class) model that we originally worked from. In such case the split will be 79.90% : 20.10%. 79.90% of the data will be used as a training set and 20.10% of the data will be used as a testing set (Please see figure 1.9

– Partition). When we separate data into training and testing sets we can minimize the discrepancies effect and understand clearly the characteristics of the model and simulate similar larger samples for a better indication in the future. The Class\_without\_reviews nugget created using the training set, will be validated by the test set. The testing set using the training set will predict the model accuracy and place confidence levels for each comment. We need to perform the same action again using the Class with reviews model nugget, which has already allocated the reviews in categories to be validated by the testing set. In the final round of testing, a comparison is made between the Class\_without\_reviews (comments not categorized) nugget and the Class\_with\_reviews nugget (categorized) to come up with the final results.

### **SPSS modeler results**

The results of the final model output (Class\_with\_reviews) will focus on 3 different model nuggets (diamond shaped) called the Class (Actual data and model prediction), Class without reviews (using the training set and validation set of actual data), Class with reviews (using the training set and validation set of actual data). Please see figure 1.91 – Full model diagram.

### **Class nugget results**

The class nugget contains 100% of the raw data that is analyzed in terms of finding the number of false and true review comments in the class field (target) based on 13 predictors (Helpful comment, Short Message, Quality, Size, Smell, Color, Love it, Like it, Problem, Gift, Delivery). Running the Class nugget, will provide us with the most important predictor in terms of a false or a true comment based on my prediction in the (Class field), the model prediction output (\$R-Class) and the confidence level in this model prediction (\$RC-Class). The most important predictor in the Class nugget was short message with (28%), helpful comment (22%), Color (20%), Love it (20%), Size (20%) and Gift (5%). Taking the most important predictor as short message the following rules are predicted:

- If this comment is not short and does not contain the word color and the word gift, then its true.

- If the comment is not short and does not contain the word color but contains the word gift, then it is most likely false.
- If the comment is not short and contains the word color, then it is true.
- If it is a short comment and does not contain a helpful comment, or the comment is missing without the word love then it is a true comment.
- If the comment is short and does not contain a helpful comment, or the comment is missing, but includes the word love and does not include the word size then the comment is true.
- If the comment is short and does not contain a helpful comment or the comment is missing but includes the word love and size then the comment is false. Please see appendix figure 2.0 – Class predictor importance).

The analysis node evaluated the ability of the model to generate accurate predictions comparing the Class column filled by my internal gut feeling when assigning false and true to the comments with the prediction from the model itself named \$R-Class. The output was 735 comments categorized as correct (73.5%) and 265 comments (26.5%) out of the total 1,000 reviews were wrong, meaning (those that were allocated false by me were predicted as true by the model and vice versa. Please see appendix figure 2.1 – Class analysis).

#### **Class without reviews nugget results**

This nugget does not involve reviews being allocated to special categorization as we will do when we test for Class with reviews model but involves partitioning the data using the training set of 79.90% (799 rows of training data was used) and 20.1% (201 rows of testing data was used). This nugget results in the following output:

- If it is not a short comment, does not contain the word gift or the word color then the review is true.
- If it is not a short comment, does not contain the word gift, but contains the word color then it is true.



- If its not a short comment and contains the word gift then its false.
- Identifying a comment as short is considered a true comment.

The most important predictor in the Class without review model are: gift (67%), short message (19%) and color (14%). Please see appendix figure 2.3 – Class without reviews predictor importance.

If we check the prediction of this model with 79.90% of the data, we get 73.34% (586) correct reviews and 26.66% (213) as wrong reviews. Testing the model with the 20.1% of data yields 72.14% (145) correct reviews and 27.86% (56) wrong reviews. Please see appendix figure 2.4 – Class without reviews analysis.

### Reviews Nugget results

Using the data text analytics node and analyzing its contents creates categories and allocates each comment/review to a category, so we can depict later which of these categories were the most predictive in terms of a false or true review and the effect it has on the predicted model in terms of a higher or lower prediction percentage. Each of these categories have sub categories under them as well, which forms a word library. The categories created for allocating the comments were based on the following: Anatomy, Arts, Astronomy, Clothes, Color, Commercial Establishment, Consumer Electronics, Day, Deal, Design and Crafts, Picture, Extras, Family Structure, Fasteners, Finance, Gift, House rooms, Implements, Look, Makeup, Material, Odor, Ordering, Person, Pockets, Product Attributes, Purse, Size, Sports, Strategic Consulting. (Please see appendix figure 2.5 – Review Categories).

### Class with Reviews Nugget results

According to the review/comments results and their allocation to each category described above, the most important predictor to classify true/false reviews or comments were based under the Consumer category with (18%) importance, pockets (17%), Makeup with subcategory smell (14%), Product attributes (12%), Astronomy with stars sub category (10%), Purse (9%), Zippers and Leather (7%) and Clothing (5%). (Please see appendix figure 2.6 – Class with review predictor

importance). If we take a sample comment from the consumer category, which has the highest predictor importance of 18%, we can deduce that the comment contained words which placed it in this high category predictor. The message in row 30 says “The only reason I gave this a 4-**star** rating was because of the cross-body **strap**. I wish it was wide as the **shoulder strap** and perhaps thicker considering it drags more when you just let it hang, especially since one of the main reasons I’ll be using that feature is when I am carrying heavier **items**. The **straps** just ends up digging into my **shoulder** uncomfortably. Otherwise, I love the size, the **pockets**, the **look** and since the bag I bought was sea green, the **gold hardware** blended very well for an overall **nice** coloring. No regrets, would purchase again, especially if they fix the width of the cross-body **strap**”. Any comment which contains (gold, hardware, item, mobile, phone, smartphone, computer, tablet, kindle, body strap, color, nice color, consumer, fasteners, look, pockets, accessories, shoulder, star) all contributed to the comment being placed in the consumer category. Prediction of the model points out that these types of reviews are false ones with a confidence in the prediction of 0.6 or 60%. However, if these reviews in the consumer category contained extra words such as clothes, handbag, bag, finance, purchase, room, odor, smell, feel of leather, leather, makeup then the message will get an extra boost of being a true message and the model prediction confidence for that message is 0.946 or (94.6%). The following comment is an example which contains some of the words with a high confidence level. The message in row 33 says “ I really like my new **handbag** it is large enough to carry everything I need plus I have more **room** than what I thought I would. It has a very soft **leather feel** to it. Only drawback is the inside has a **odor**. Before I bought his **handbag**, I read some of the remarks and saw that was a problem with a few people. I thought I would buy it anyway. So far, I really do like my new **purchase**. I have had this for a week and hoping the **smell** will eventually come out soon”. (Please see appendix figure 2.7a – Messages in category, 2.7b -Message category confidence - message 30) and 2.7c – Message category confidence - message 33).

### **Comparison and success rate for model prediction results**

To understand if the model created (Class with reviews) was successful, we need to compare the two models, Class without reviews and Class with reviews, to see the change in success rate that took place when reviews were added and categorized in the analysis. The results show that adding the reviews and categorizing them gave us an extra 1% increase in success rate in predicting the model using the training set and 0.99% using the testing set. (Please see appendix figure 2.8 – Success rate). Although this may seem a small number, but a 1% change in success rate is a good measure that such model can be used to predict larger numbers of data (ie., more than 1,000) with a higher success rate if more fields were added in terms of data column attributes.

### **Challenging aspects and how to do better the next time**

#### **Python commands**

SPSS modeler is created for those that don't have an extensive background in Python language. However, the code that was involved in automating the process of testing and executing the stream output, was based on the number of nodes used. In this analysis, there were 7 different types of nodes used (Analysis, CHAID, Distribution, Data Audit, Table, Partition) to create the whole module. It was challenging to understand the stream automation script code method and how to use it, since without it the model could not have been executed and 100's of code lines was needed to produce the outputs. In this regard, learning python code automation is a must (ongoing process). (Please see appendix figure 2.9a – Python part1 & 2.9b – Python part2 & 2.9c - Python part3).

#### **Web Scraping**

Knowledge in web scraper was required in terms of creating the site map for the Scarleton large drawstring handbag. Understanding where to place the selectors and what type of selectors to choose to come up with the site map was a challenging aspect. I also had difficulties installing it for the first time on my Mac as it produced errors in installation. I had to re-install Google browser again and re-install the web scraper plug-in and it worked perfectly. This process took almost 2

weeks to grasp the idea of Google web scraping and extraction of the data. (Please see appendix figure 3.0 – Amazon scraping sitemap).

### **Classifying data (True/False)**

This was the most time consuming of all the required items before SPSS analysis. Once the data was extracted it had to be organized into fields, separated into columns and cleaned in terms of data that is not required such as page links, where the pagination came from and other filler information that does not add to the analysis output. 1,000 comments were read, analyzed and allocated into the 13 attributes, which are explained in the steps in the Google analysis section of this paper. Using pivot tables could have eased attribute identification and reduced time in finding the information in the excel data sheet, however, I only came to know about pivot tables when it was later introduced in our Business Intelligence class.

### **SPSS syntax errors**

Using SPSS modeler version 18 was not an easy task, since it was the first time for me to get acquainted with it. Adding to that, the classification module, a core basis of this analysis, was not introduced to us in the Business Intelligence class at an earlier stage. As such, I had to learn by taking online courses on YouTube, reading the IBM scripting guide and asking for the BI professor help.

### **Human prediction errors**

Classifying the data requires human interaction at in the first stage of classifying the data into 1's (true) and 0's (false) reviews. In this scenario a human may classify a true review as false and vice versa. If I had the opportunity to use machine learning (human computer interaction software along with SPSS (out of this project scope), a higher success factor of more than 1% could have been achieved.

### **Conclusion/Recommendations**

It is not an easy task to identify fake reviews as they are pieces of information that are describing products/services based on a customer visit/usage of a product. However, listening to the language

i.e., very enthusiastic, very negative or all in CAPS reviews can pose as a flag. If comments seem to look like manuals, offer marketing jargon and very technical in nature, then it's a warning sign that they are fake. When someone explains why they don't like something, then this is a good sign that the review is not a fake due to the attention given in the review detail. Since, the output for the class with reviews yielded the best success rate, companies need to look at categorizing the review word content when generating a model and of course this model should be catered to the nature of their business. In such a case, these identified words will be placed in a word library (categorized words) to detect fallacies in comments, when they are screened in artificial intelligence software.

### **Suggestions for Online retail companies**

1. Add mood ticker boxes so when a comment is written, customers can tick on "good mood, bad mood, not sure, mixed feelings" to give an idea on customer sentiment.
2. An algorithm filter should connect purchaser with review made. An absence of a certified purchase is a clue that it is not genuine.
3. Anonymous users should not be allowed to place reviews on a website/app.
4. A review level (dynamic graphic) from green to red can be used while the reviewer is writing. Feeling of review importance will encourage a non-fake review.
5. Having rules/policies (Terms of Service) will help customer understand the importance of a review. Placing responsibility on the customer, will insure that the reviews are genuine.
6. Adding a voting value associated with product review. Higher reliability voting percentages are always better in terms of indicating how well the review affected a particular purchase. Example, if a person purchases a product based on a product review (found under helpful comment on Amazon/online retail pages) and finds that the review was deceptive upon receiving the product, those influential reviews should be connected to the purchaser's product purchase, where he/she got the information from and have them flagged by the purchaser as misleading reviews/comments the next time someone wants to gain an idea on the same product.

### **Project success factor**

My business intelligence professor, after reviewing the approach I took into coming up with the best alternative model (Class with Reviews) and seeing the output of the model himself, asked if I can present it to the class to show them my work. With the limited resources I had, this project can be scaled further, whereby human intelligence software can be used along with SPSS to simulate my model on larger data sets.

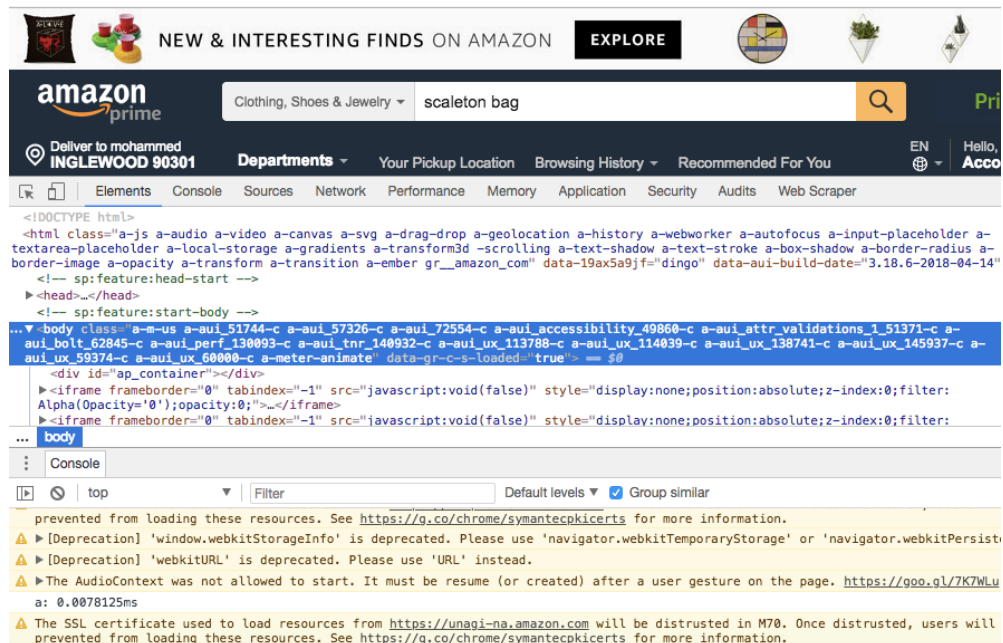
Thank you for giving me the opportunity to share my business intelligence work in Management 647, this project will be enhanced further as I am looking forward in researching it more aggressively in my Ph.D. education, which I am embarking for after my M.BA graduation in Fall 2018. The research will focus on predicting a model similar to this one, but with a higher success rate, using human interaction software, behavioral and sentimental analysis basing it on the outcome of a written review when in a particular mood set.

## Appendix

### KPI's (1.0)

Project Key Performance Indicators				
Strategic Initiative	KPI	Weight	My score	Why this score?
Improve reporting and transparency	Categorization of 1,000 amazon comments/reviews	20%	15%	True/False classification in the Class column is not 100% correct (human error)
Increase reliable product/service reviews	Effective model creation (atleast 80% accuracy)	35%	30%	Model was created successfully (75% accuracy in model prediction from data). I placed 80% success rate - could be a harsh level of accuracy, but with the limited resources I had, this was not a bad output
Improve model development	1% increase in Success rate comparing class with reviews and class without reviews	5%	5%	Achieved the 1% success rate I applied to my model when comparing class without reviews to class with reviews.
Improve technical skills	Python language knowledge	20%	15%	Pyhton language requires more time to understand it. I believe that I was able to understand what the project scope required and was able to execute the nodes successfully.
Improve certain skill	Google web scraping knowledge	20%	20%	Able to completely extract the information after obtaining training on Youtube. Data extracted was used for the model predictions without any issues.
		100%	85%	

### Google web scraping (1.1)



The image shows the Amazon homepage with a web scraper overlay. The scraper displays the HTML structure of the page, including the header, navigation bar, and product listings. The scraper also shows the console output, which includes warnings about deprecated APIs and SSL certificate issues.

NEW & INTERESTING FINDS ON AMAZON EXPLORE

amazon prime

Clothing, Shoes & Jewelry skeleton bag

Deliver to mohammed INGLEWOOD 90301

Departments Your Pickup Location Browsing History Recommended For You

Elements Console Sources Network Performance Memory Application Security Audits Web Scraper

```
<!DOCTYPE html>
<html class="a-js a-audio a-video a-canvas a-svg a-drag-drop a-geolocation a-history a-webworker a-autofocus a-input-placeholder a-textarea-placeholder a-local-storage a-gradients a-transform3d -scrolling a-text-shadow a-text-stroke a-box-shadow a-border-radius a-border-image a-opacity a-transform a-transition a-ember gr_amazon_com" data-19ax5a9jf="dingo" data-aui-build-date="3.18.6-2018-04-14">
  <!-- sp:feature:head-start -->
  <head>
    <!-- sp:feature:start-body -->
    ... body class="a-m-us a-aui_51744-c a-aui_57326-c a-aui_72554-c a-aui_accessibility_49860-c a-aui_attr_validations_1_51371-c a-aui_bolt_62845-c a-aui_perf_130093-c a-aui_tnr_140932-c a-aui_ux_113788-c a-aui_ux_114039-c a-aui_ux_138741-c a-aui_ux_145937-c a-aui_ux_59374-c a-aui_ux_60000-c a-meter-animate" data-gr-c-s-loaded="true" == $0
    <div id="ap_container">
      <iframe frameborder="0" tabindex="1" src="javascript:void(false)" style="display:none;position:absolute;z-index:0;filter:Alpha(Opacity='0');opacity:0;"></iframe>
      <iframe frameborder="0" tabindex="1" src="javascript:void(false)" style="display:none;position:absolute;z-index:0;filter:
    ... body
  </div>
  </html>
```

prevented from loading these resources. See <https://g.co/chrome/symantecpkicerts> for more information.

[Deprecation] 'window.webkitStorageInfo' is deprecated. Please use 'navigator.webkitTemporaryStorage' or 'navigator.webkitPersist'.

[Deprecation] 'webkitURL' is deprecated. Please use 'URL' instead.

The AudioContext was not allowed to start. It must be resume (or created) after a user gesture on the page. <https://goo.gl/7K7WLu>

a: 0.0078125ms

The SSL certificate used to load resources from <https://unagi-na.amazon.com> will be distrusted in M70. Once distrusted, users will be prevented from loading these resources. See <https://g.co/chrome/symantecpkicerts> for more information.

### Web Scrapping code document



Webscrapping\_code.docx

## Amazon scraping sheet (1.2)

ID	Reviews	Helpful comment	Short message	Long message	Quality	Size	Smell	Color	Love it	Like it	Problem	Gift	Delivery	Class
15207589512352	Well pleased with this handbag. ... it's good quality ... just the right size and the color is awesome...	0	1	0	1	1	0	1	0	0	0	0	0	T
15207587381520	Great purse! I have gotten many compliments and it has lasted longer than any of my other purses!	0	1	0	0	0	0	0	0	0	0	0	0	T
15207586691202	It was a little too large for me, but nice. The color was a true blue not navy. Nice style Bought this bag for my daughter. Looks great. You can't tell it's not leather. Picked out a brown one and she loved the color. I liked the bag so much I ordered a burgundy one.	0	1	0	0	0	0	1	0	0	0	0	0	T
15207586521128	It has a lot of space for those of us who take our whole house in our purse. I use this as my go to purse for my lengthy work commute. I can even stick a small lunch bag in there. I can't decide if it looks cheap or not but it was certainly cheap haha. Anyway, I wanted something that wasn't leather and could take a beating on the bus/shuttle rides I use to commute and it has definitely held up.	0	0	1	0	0	0	0	0	0	0	0	0	T
15207590982882	This purse was exactly what I wanted. It is a large bag that can fit all my things and yet comfy to wear when it gets heavy. Good amount of pockets and it looks great! Lots of compliments. Love this bag!	0	1	0	0	0	0	0	0	0	0	0	0	F
15207589272256	Really beautiful, stylish bag! Just love!	0	1	0	0	0	0	0	0	0	0	0	0	F
1520758618980	Doesn't really look like real leather, but it's cute. Maybe I expected too much. I'll use it but don't plan on buying any more. Save your money and go to designer handbag outlet stores	0	1	0	0	0	0	0	0	1	0	0	0	T
15207589272257	This is my first bag by Scarleton and it definitely won't be my last! I will be ordering more colors of this bag. I have another, different style, bag by Scarleton already ordered, but I haven't received it yet. I can tell it will be a winner though. If you didn't know what this bag was made from, you would be hard pressed to say it wasn't leather. The material is sturdy and thick and substantial. The hardware is really pretty and looks like it will last for a very long time. The zippers work smoothly. The bag is the absolute perfect size. Not too huge, but certainly not small by any means. I love the crossbody strap that allows me to carry the bag hands free. This is my new favorite brand!	0	0	1	0	1	0	1	1	0	0	0	0	T
15207587791701	Recommend!	0	1	0	0	0	0	0	0	0	0	0	0	F
15207587791701	Recommend!	0	1	0	0	1	0	0	1	0	0	0	0	T

## SPSS excel node (1.3)

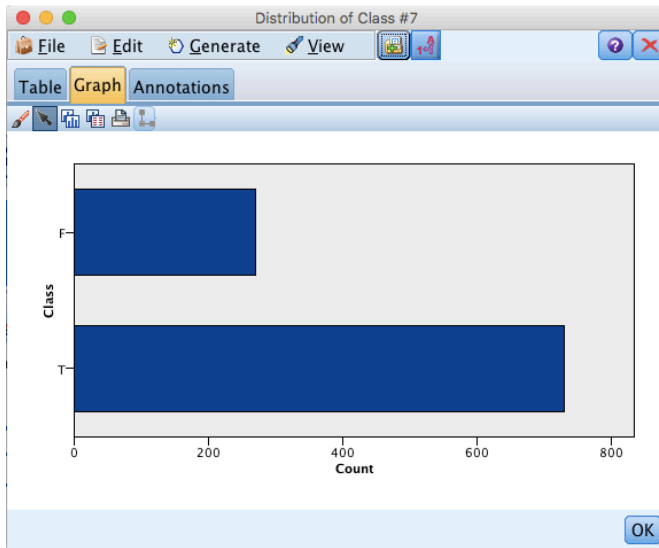
ID	Reviews	Helpful comment	Short message	Long message	Quality	Size	Smell	Color	Love it	Like it	Problem	Gift	Delivery	Class
15207589512352	Well pleased with this handbag. ... it's good quality ... just the right size and the color is awesome...	0.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	T
15207587381520	Great purse! I have gotten many compliments and it has lasted longer than any of my other purses!	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	T
15207586691202	It was a little too large for me, but nice. The color was a true blue not navy. Nice style Bought this bag for my daughter. Looks great. You can't tell it's not leather. Picked out a brown one and she loved the color. I liked the bag so much I ordered a burgundy one.	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	T
15207586521128	It has a lot of space for those of us who take our whole house in our purse. I use this as my go to purse for my lengthy work commute. I can even stick a small lunch bag in there. I can't decide if it looks cheap or not but it was certainly cheap haha. Anyway, I wanted something that wasn't leather and could take a beating on the bus/shuttle rides I use to commute and it has definitely held up.	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	T
15207590982882	This purse was exactly what I wanted. It is a large bag that can fit all my things and yet comfy to wear when it gets heavy. Good amount of pockets and it looks great! Lots of compliments. Love this bag!	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	F
15207589272256	Really beautiful, stylish bag! Just love!	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	F
1520758618980	Doesn't really look like real leather, but it's cute. Maybe I expected too much. I'll use it but don't plan on buying any more. Save your money and go to designer handbag outlet stores	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	T
15207589272257	This is my first bag by Scarleton and it definitely won't be my last! I will be ordering more colors of this bag. I have another, different style, bag by Scarleton already ordered, but I haven't received it yet. I can tell it will be a winner though. If you didn't know what this bag was made from, you would be hard pressed to say it wasn't leather. The material is sturdy and thick and substantial. The hardware is really pretty and looks like it will last for a very long time. The zippers work smoothly. The bag is the absolute perfect size. Not too huge, but certainly not small by any means. I love the crossbody strap that allows me to carry the bag hands free. This is my new favorite brand!	0.000	0.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	T
15207587791701	Recommend!	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	F
15207587791701	Recommend!	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	F

## Type node (1.4)

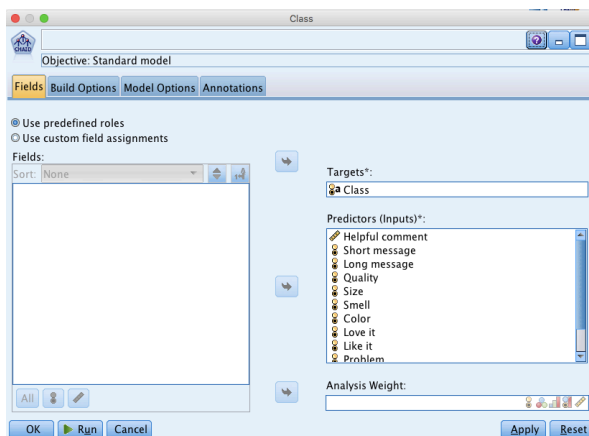
Field	Measurement	Values	Missing	Check	Role
ID	Typeless			None	Record ID
Reviews	Typeless			None	None
Helpful comment	Continuous	[0.0,204.0]		None	Input
Short message	Flag	1.0/0.0		None	Input
Long message	Flag	1.0/0.0		None	Input
Quality	Flag	1.0/0.0		None	Input
Size	Flag	1.0/0.0		None	Input
Smell	Flag	1.0/0.0		None	Input
Color	Flag	1.0/0.0		None	Input
Love it	Flag	1.0/0.0		None	Input
Like it	Flag	1.0/0.0		None	Input
Problem	Flag	1.0/0.0		None	Input
Gift	Flag	1.0/0.0		None	Input
Delivery	Flag	1.0/0.0		None	Input
Class	Flag	T/F		None	Target



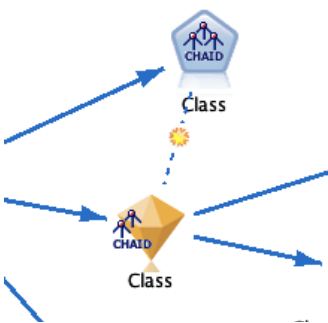
## Distribution class (1.5)



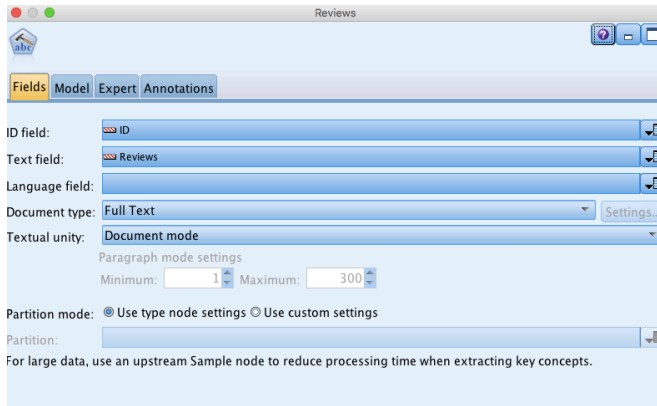
## CHAID (1.6)



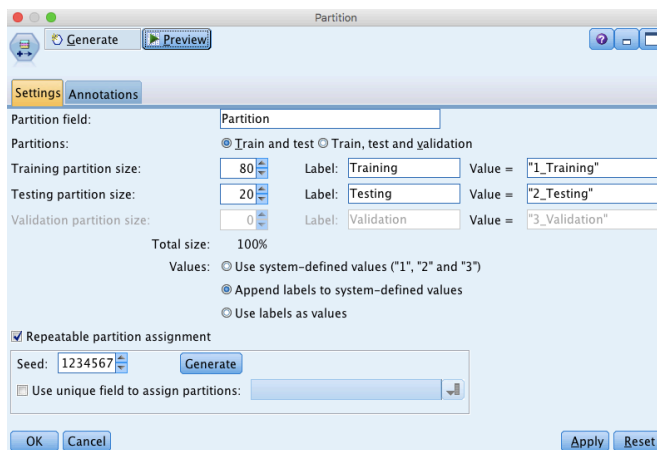
## CHAID class tree model (1.7)



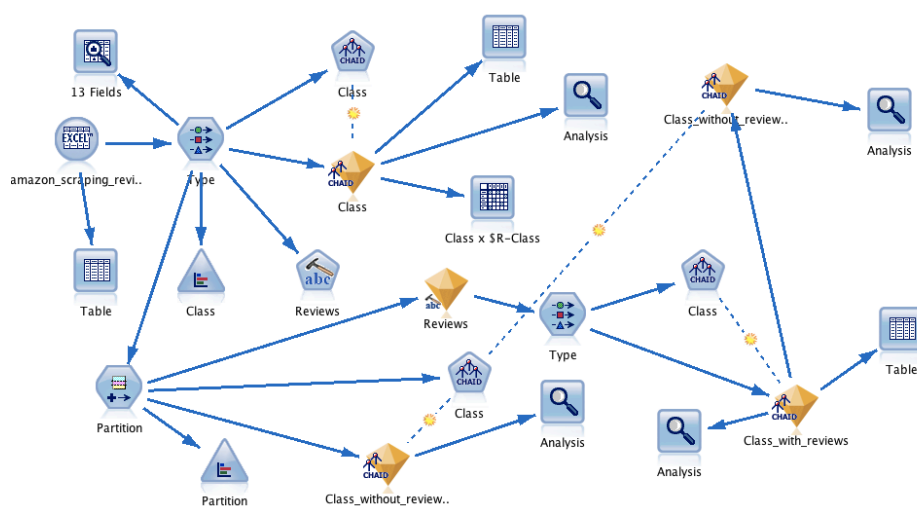
## Text mining node (1.8)



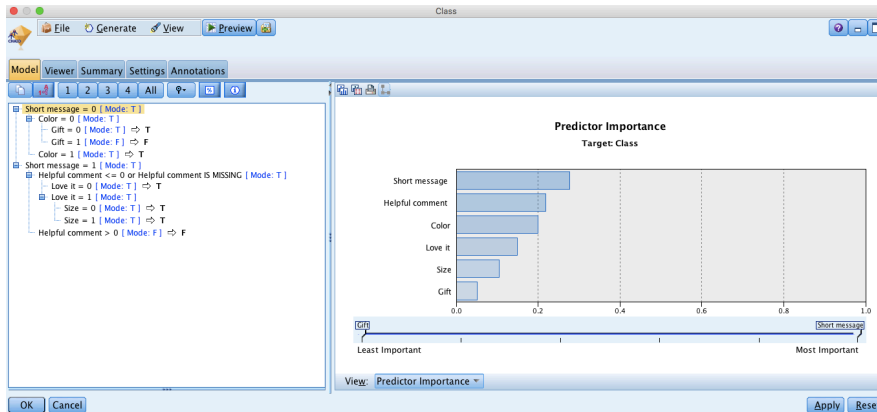
## Partition (1.9)



Full model diagram (1.91)



## Class predictor importance (2.0)



## Class analysis (2.1)

Analysis of [Class] #33

Results for output field Class

Comparing SR-Class with Class

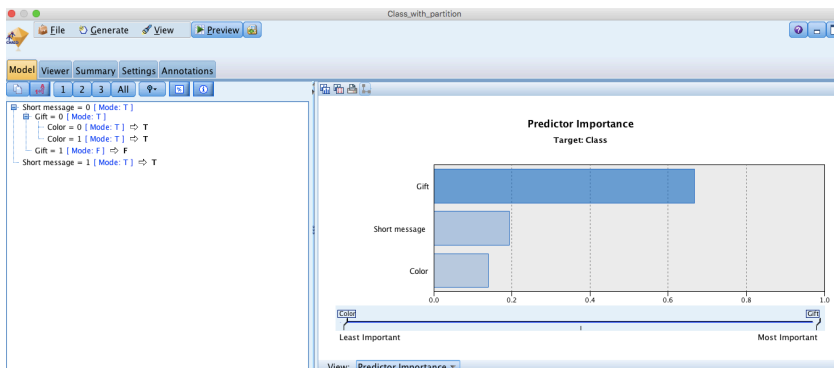
	Correct	Wrong	Total
Count	735	265	1,000
Percentage	73.5%	26.5%	

## Distribution partition (2.2)

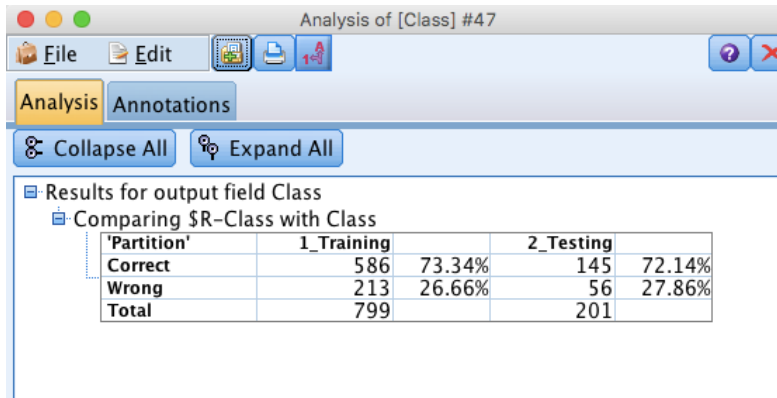
Distribution of Partition #2

Value	Proportion	%	Count
1_Training	79.9	79.9	799
2_Testing	20.1	20.1	201

## Class without reviews predictor importance (2.3)



## Class without reviews analysis (2.4)



Analysis of [Class] #47

File Edit ?

Analysis Annotations

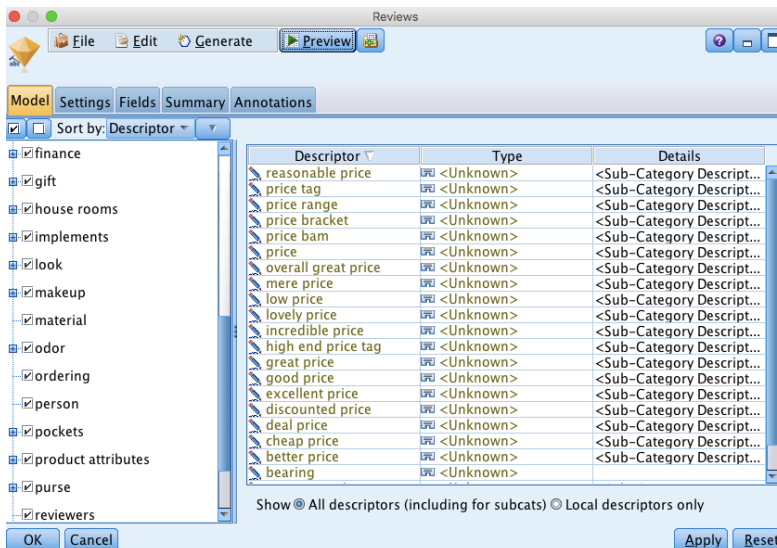
Collapse All Expand All

Results for output field Class

Comparing \$R-Class with Class

'Partition'	1_Training	2_Testing
Correct	586 73.34%	145 72.14%
Wrong	213 26.66%	56 27.86%
Total	799	201

## Review categories (2.5)



Reviews

File Edit Generate Preview ?

Model Settings Fields Summary Annotations

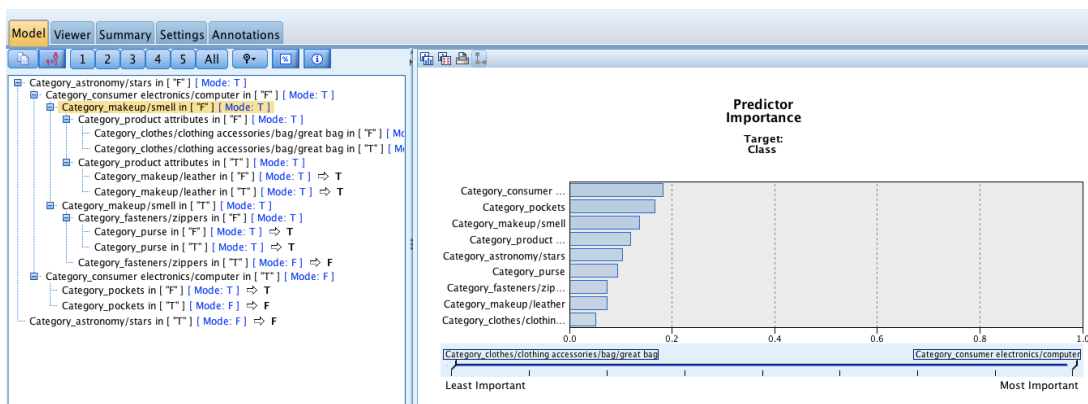
Sort by: Descriptor

Descriptor	Type	Details
reasonable price	<Unknown>	<Sub-Category Descript...
price tag	<Unknown>	<Sub-Category Descript...
price range	<Unknown>	<Sub-Category Descript...
price bracket	<Unknown>	<Sub-Category Descript...
price bam	<Unknown>	<Sub-Category Descript...
price	<Unknown>	<Sub-Category Descript...
overall great price	<Unknown>	<Sub-Category Descript...
mere price	<Unknown>	<Sub-Category Descript...
low price	<Unknown>	<Sub-Category Descript...
lovely price	<Unknown>	<Sub-Category Descript...
incredible price	<Unknown>	<Sub-Category Descript...
high end price tag	<Unknown>	<Sub-Category Descript...
great price	<Unknown>	<Sub-Category Descript...
good price	<Unknown>	<Sub-Category Descript...
excellent price	<Unknown>	<Sub-Category Descript...
discounted price	<Unknown>	<Sub-Category Descript...
deal price	<Unknown>	<Sub-Category Descript...
cheap price	<Unknown>	<Sub-Category Descript...
better price	<Unknown>	<Sub-Category Descript...
bearing	<Unknown>	<Sub-Category Descript...

Show All descriptors (including for subcats) Local descriptors only

OK Cancel Apply Reset

## Class with review predictor importance (2.6)



### Messages in category (2.7a)

30	15207590212617.000	The only reason I gave this a four star rating was because of the cross-body strap. I wish it was as wide as the shoulder strap and ...
31	15207588441946.000	I just received my handbag today and I love it! It came on the scheduled date. Other reviews said it had a bad odor, however, mine ...
32	15207591032907.000	I love this bag. I bought it in black. The style is exactly what I've been searching for. It didn't have the center pocket. There was n...
33	15207590252633.000	I really like my new handbag it is large enough to carry everything I need plus I have more room than what I thought I would. It has...

### Message category confidence - message 30 (2.7b)

	ategic consulting	Category_strategic consulting/price	Category_strategic consulting/price/great price	Category_strategic consulting/price/price tag	\$R-Class	\$RC-Class
21	F	F	F	F	T	0.713
22	F	F	F	F	T	0.713
23	F	F	F	F	T	0.783
24	F	F	F	F	T	0.713
25	F	F	F	F	T	0.713
26	F	F	F	F	T	0.713
27	T	F	F	F	T	0.713
28	F	F	F	F	T	0.713
29	F	F	F	F	T	0.713
30	F	F	F	F	F	0.600
31	T	F	F	F	T	0.783
32	T	T	T	F	F	0.800
33	F	F	F	F	T	0.946

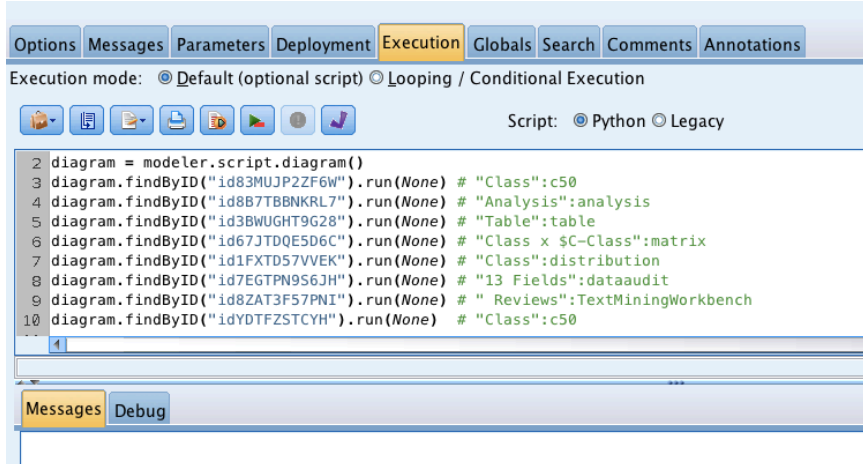
### Message category confidence - message 33 (2.7c)

21	F	F	F	T	0.713
22	F	F	F	T	0.713
23	F	F	F	T	0.783
24	F	F	F	T	0.713
25	F	F	F	T	0.713
26	F	F	F	T	0.713
27	T	F	F	T	0.713
28	F	F	F	T	0.713
29	F	F	F	T	0.713
30	F	F	F	F	0.600
31	T	F	F	T	0.783
32	T	T	T	F	0.800
33	F	F	F	T	0.946

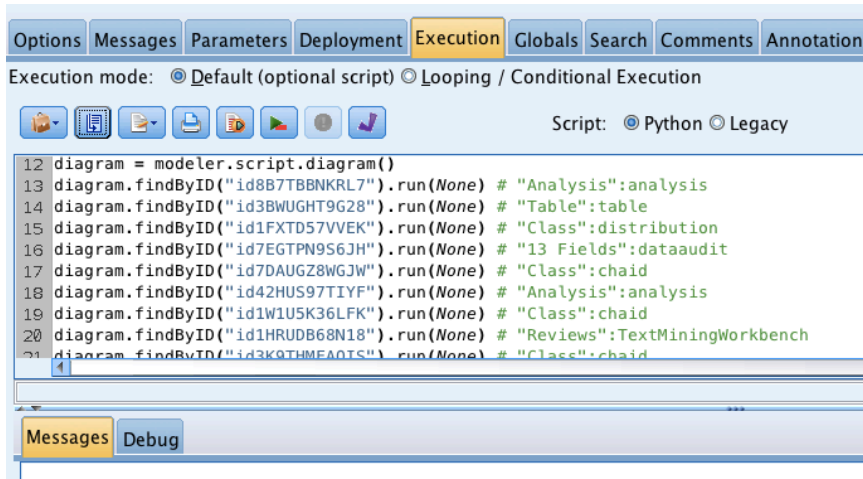
### Success rate (2.8)

SPSS Modeler					
Analysis			Analysis		
Annotations			Annotations		
Collapse All Expand All			Collapse All Expand All		
Results for output field Class			Results for output field Class		
Comparing \$R-Class with Class			Comparing \$R-Class with Class		
'Partition'	1_Training	2_Testing	'Partition'	1_Training	2_Testing
Correct	586 73.34%	145 72.14%	Correct	594 74.34%	147 73.13%
Wrong	213 26.66%	56 27.86%	Wrong	205 25.66%	54 26.87%
Total	799	201	Total	799	201

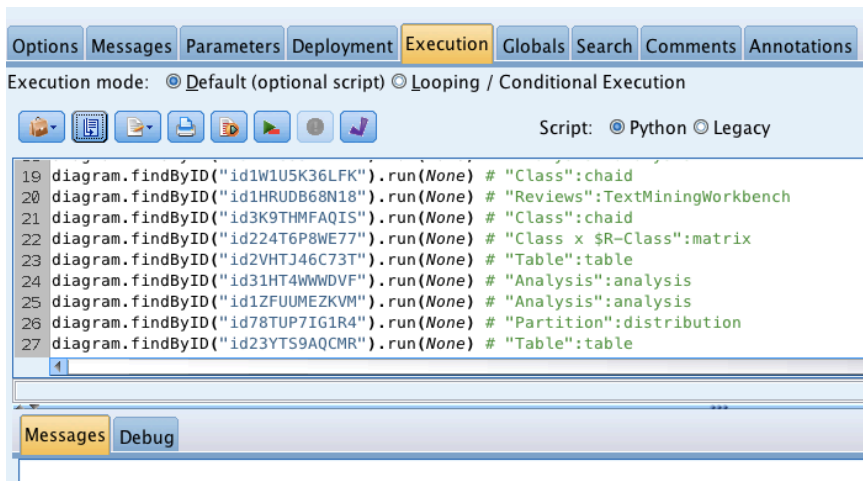
## Python part 1 (2.9a)



## Python part 2 (2.9b)



## Python part 3 (2.9c)



### Amazon scraping site map (3.0)

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4 <script type="text/javascript">window.NREUM||(NREUM={});NREUM.info={"beacon":"bam.nr-
  data.net","errorBeacon":"bam.nr-
  data.net","licenseKey":"f595f29c1c","applicationID":"5475963","transactionName":"IA1fFUUNWQ1TFB1BRgI
  BWhiYEV0OQQ==","queueTime":7782,"applicationTime":891,"agent":""}</script>
  <script type="text/javascript">window.NREUM||(NREUM={});NREUM.require=function(a){return function(n){
```



## **BIBLIOGRAPHY**

Business Analytics Software, IBM SPSS Data Mining Workshop, Commercial Edition, March 2012

IBM SPSS Modeler 18.0 Algorithms Guide, 2016

Sharda, Delsen et al., Business Intelligence, A Managerial Perspective on Analytics, 3<sup>rd</sup> edition, 2014

IBM Knowledge Center, retrieved from <https://www.ibm.com/support/knowledgecenter/en>

Kelleher, Mac Namee, et al., Fundamentals of machine learning for predictive data analytics, Algorithms, Worked examples, and Case Studies, MIT Press, 2015