

ASSOCIATION RULE MINING WITH TWEETS: THINKING OUTSIDE
THE BASKET

A THESIS SUBMITTED TO
THE FACULTY OF ARCHITECTURE AND ENGINEERING
OF
EPOKA UNIVERSITY

BY

IGLI TOLA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JULY, 2021

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled “**Association Rule Mining with Tweets, Thinking Outside the Basket**” and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Dr. Arban Uka
Head of Department
Date: July, 15, 2021

Examining Committee Members:

Dr. M. Maaruf Ali (Computer Engineering)

Dr. Iglı Hakrama (Computer Engineering)

Dr. Mirela Alhasani (Computer Engineering)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Igli Tola

Signature: _____

ABSTRACT

ASSOCIATION RULE MINING WITH TWEETS: THINKING OUTSIDE THE BASKET

Tola, Igli

M.Sc., Department of Computer Engineering

Supervisor: Dr. Maaruf Ali

In the past few years, the usage of social media around the world has impacted people's lifestyles because of the widespread use of recommender systems. As Twitter being of the main social media platforms, a high interest on text mining for this social media platform has been observed. Because of this, the need of exploring relationships among words in Twitter data is of a high importance. The main purpose of this thesis is to observe existing association rules mining techniques and use those techniques to build correlations between words in tweets. Market Basket example is the most common example used in association rule mining. By using the logic of market basket practice, this study aims to use a similar approach to build a logical file composed of tweets to understand the associations between tweets. This thesis allowed me to understand better on the steps needed to take, in order to have a well formatted file containing the tweets by using R programming. Results show us that different words between different tweets have a high measure of Lift, making the association rules between words meaningful.

Keywords: *Association Rules, Apriori Algorithm, Tweets, R coding, Market Basket*

ABSTRAKT

RREGULLAT E ASOCIACIONIT ME TWEETS SIPAS SHEMBULLIT TË SHPORTËS SË TREGUT

Tola, Igli

Master Shkencor, Departamenti i Inxhinierisë Kompjuterike

Udhëheqësi: Dr. M. Maaruf Ali

Gjatë viteve të fundit, përdorimi i rrjeteve sociale përreth gjithë botës ka ndikuar në mënyrën e jetesës së njerëzve për shkak të përhapjes së madhe të sistemeve të rekomandimeve. Duke qënë se Twitteri është një nga rrjetet sociale kryesore, është vëzhguar një interes shumë i madh në minimin e tekstit për këtë rrjet social. Për shkak të kësaj, nevoja për të exploruar lidhjet midis fjalëve të ndryshme në të dhënat e Twitter është me rëndësi të madhe. Qëllimi kryesor i kësaj teze është për të vëzhguar rregulla asociimi, dhe ato teknika të ndërtojnë lidhje midis fjalësh në “tweet-e”. Shembulli i “Shportës së Tregut” është shembulli më i shpeshtë që përdoret në minimin e rregullave të asociimit. Duke përdorur logjikën e “Shportës së Tregut”, ky studim synon që të përdori një qasje të ngjashme, për të ndërtuar një dosje me “tweet-e” për të kuptuar lidhjet midis “tweet-save”. Kjo tezë më ndihmoi të kuptoja më mirë hapat që duhet të ndiqen në mënyrë që të kemi një dosje të formatuar mirë që përmban “tweet-e” duke përdorur gjuhën e programimit R. Rezultatet tregojnë që fjalë të ndryshme mes “tweet-eve” kanë një vlerë të madhe të njësisë matese “Lift”, duke i bërë rregullat e asociimit me kuptimplotë.

Fjalët kyçe: Rregullat e Asociimit, Algoritmi Apriori, Tweets, Programim R

Dedicated to my wonderful parents, my brother, and my girlfriend.

TABLE OF CONTENTS

ABSTRACT.....	iv
ABSTRAKT	v
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF EQUATIONS	xi
LIST OF ABBREVIATIONS.....	xii
CHAPTER 1	1
INTRODUCTION	1
1.1 Importance of Social Media	1
1.2 Thesis Objective	2
1.3 Organization of the thesis.....	2
CHAPTER 2	4
LITERATURE REVIEW	4
2.1 Importance of Machine Learning.....	4
2.1.1 Supervised Machine Learning	6
2.1.2 Semi - Supervised Machine Learning.....	7
2.2 Unsupervised Machine Learning.....	7
2.2.1 Clustering.....	8
2.2.2 Anomaly Detection.....	9
2.3 Association Rules Mining	9
2.3.1 Market Basket, an Association Rule Mining Practice	10
2.3.2 Common Measures in Association Rule Mining.....	11
2.4 The Apriori Algorithm	13

2.5 R Programming	15
CHAPTER 3	18
METHODOLOGY	18
3.1 Introduction	18
3.2 Data Set Used	18
1.3 Specifications on the tweets data	19
3.4 Data Collection.....	21
3.4.1 R Twitter Options	22
3.5 Building the Transaction File.....	25
3.6 Data Cleaning	27
3.6.1 Actual Clean Up	28
3.6.2 Removing Specific Words	29
3.6.3 Cleaning with grepl	30
CHAPTER 4	33
RESULTS AND DISCUSSIONS	33
4.1 Implementing Association Rules in R.....	33
4.2 A quick Plot.....	35
CHAPTER 5	37
Conclusions and Future Work	37
5.1 Conclusions	37
5.2 Limitations and Future Work	38
REFERENCES	39

LIST OF TABLES

Table 1 Examples of Association Rules [23].....	10
Table 2 Reviewing Association Rules [23]	15
Table 3 Second way of data representation [23].....	16
Table 4 Third way of data representation [23].....	16
Table 5 How tweets should look like.....	19
Table 6 Most Frequent Items.	28

LIST OF FIGURES

Figure 1 Machine Learning Algorithms	6
Figure 2 Types of Unsupervised Learning.....	8
Figure 3 The Apriori Algorithm	14
Figure 4 Developer Twitter Credentials.....	21
Figure 5 Getting the keys and tokens.....	21
Figure 6 Twitter Keys and tokens in .txt.....	22
Figure 7 Returned list of objects.....	23
Figure 8 An example of how a tweet looks like.	24
Figure 9 Tweets into tokens opened with Excel	26
Figure 10 read.transactions in basket format	27
Figure 11 Pre-Formatted transactions data into Data Frame	29
Figure 12 Formatted Transactions Data Frame.....	31
Figure 13 Tweets as transactions, basket format.	32
Figure 14 Sorted Rules.....	34

LIST OF EQUATIONS

Equation 1 Support	11
Equation 2 Confidence.....	11
Equation 3 Lift	12

LIST OF ABBREVIATIONS

IoT	Internet of Things
ML	Machine Learning
RHS	Right Hand Side
LHS	Left Hand Side
ARM	Association Rule Mining
Supp	Support
Conf	Confidence
API	Application Programming Interface
DF	Data Frame
TID	Transaction ID

CHAPTER 1

INTRODUCTION

This thesis consists of four chapters, where every chapter will explain the whole scope of association rules importance in datasets, and how beneficial it is for the use of twitter data. Chapter one gives us an overview of the importance of Twitter data, motivation for this study and the overall structure will be discussed.

1.1 Importance of Social Media

Information Technology and computer networks has had a humongous advancement, especially in the past two decades, and this has led to an abundance of data which can be used for decision making from many different commercial and marketing organizations all over the world [1]. Due to its high number of users all over the world, using social media through different platforms such as Facebook, Instagram, Twitter, LinkedIn etc., basically everyone that has access to a smartphone that can be connected to the internet, can be a part of this virtual network. If we're talking numbers, during 2020 there was a total of 3.6 billion social network users worldwide, while it is estimated that by 2025, this number is to be 4.41 billion social network users [2] and this trend is probably going to go up in the next 10 years too. There are 500 million tweets sent each day, & that's 6,000 tweets each second. With the increasing of social media users worldwide, raw data coming from the content of these social media platforms are going to increase accordingly as well. It is a known fact that businesses see interests of social media users in order to advertise their products to the right audience. In fact, 65.8% of US companies with 100+ employees use Twitter for marketing [3]. With an ever-growing amount of social network users and the respective data that is accompanied by it, the need to understand this data and do different kinds of research on them is on high demand.

1.2 Thesis Objective

Text Analysis has a main task of extracting the text from retrieved real-world information from the web and application of text mining to visualize the text data only. Knowing that text analytics can be applied to any text data, which can be any language and containing images and emojis, making it completely unstructured data. Textual data often comes with additional challenges such as incorrect spellings, incorrect syntax of the sentences and it leads to challenges for the extraction of the correct information out of that and it's processing as well.

This raises important questions regarding text mining:

1. How can raw and unstructured data such as tweets be converted into meaningful data to be processed?
2. What machine learning technique can be efficient, when finding associations between words in a set of Tweets?

The task and the aim of this thesis is to use the market basket classic example of association rule mining. This thesis intends to use that example for tweet mining. How relevant are the items in the twitter dataset in order to see if these items (or words) are associated with each other, while making sure on retaining each tweet as its own transaction. Testing is made in R programming.

1.3 Organization of the thesis

Chapter 1: The first chapter gives us an introduction to the research area, accompanied by thesis objective, showing the aim of the research area, and also the main research questions that are observed.

Chapter 2: This chapter gives a brief explanation of Association rules. It explains how association rules works, the parameters which are needed to be taken into consideration when building our model and the problems that may occur. It also explains how

association rules are generally used in different research, and use the same idea of the methods, to build our implementation in the coming chapter.

Chapter 3: This chapter introduces the implementation of Association Rules to our Twitter data set, based on the commonly used model, the “market basket” model.

Chapter 4: In this chapter, the results that have been observed in Chapter 3 are concluded, and limitations from the implementation will be mentioned

CHAPTER 2

LITERATURE REVIEW

This chapter covers general information on what this thesis is all about. It gives information on everything we know about Association Rules and unsupervised learning, including an understanding of unsupervised learning itself, as that is the kind of algorithm that will be implemented in our thesis. We will get to know what the most common example of the use of Association Rules is and where it can also be applied to.

2.1 Importance of Machine Learning

Machine Learning is a method of data analysis which automates analytical model building. As a branch of artificial intelligence, it is based on the idea that systems can learn and observe patterns to make decisions without the help of human interactions. ML nowadays is not like it used to be in the past. How machine learning was born, is because researchers had a theory that computers could learn without being programmed to do different tasks. Early ML techniques were incapable and rigid of tolerating any variations from the training data [4]. Because of this, many researchers interested in artificial intelligence wanted to see if computers could learn from data. Nowadays, Machine Learning has improved and automated many aspects of life. It has helped a lot in the area of medicine, automating the process of diagnosing, and improving the areas of medical imaging. With no doubt, there is a huge amount of data with emerging networks, such as Internet of Things (IoT) and its billions of connected devices [5]. This fact gives ML the ability to identify unexpected patterns, and furthermore can be used to understand the processes that generate the data. But which industries are using Machine Learning? ML is used in many industries around the world, mostly the industries that have to do with large amount of data. These industries have accepted and acknowledged the importance of machine learning technology. By getting insights from these huge amounts of data,

these industries have taken advantage of ML by working more efficiently over competition.

Some of the main industries that use ML are as following [6]:

- Financial Institutions: These institutions use ML for two main reasons: to identify key insights in data to avoid fraudulent activities & also inform investors on when to trade, by identifying high risk profiles
- Government agencies: Public Safety has a lot of sources of data that can be mined for insights. These agencies have a very high need of machine learning. It can very much help in finding ways to save money and work efficiently.
- Health Care: In healthcare, the need for machine learning is very much important, as it can use data to track a patient's health in real time. This technology can help doctors in identifying and improving the diagnosis of patients so then then can proceed with treatments.
- Retail: This industry is important because data that comes from websites and/ or social media are the reason that recommendations come up while you're scrolling through your phone. Retailers often rely on machine learning to analyze data and to make the shopping experience much better by different campaigns such as marketing campaigns, price optimizations, supply planning etc. Retail data is especially important for this thesis as the example used to build our thesis is based on retail store data and the methods used on finding associations between different items that consumers purchase.
- Oil and gas: interestingly enough, machine learning is used in this industry as well to find new energy sources. It can be done by streamlining oil distribution so it can be cost-effective.
- Transportation: have you ever wondered how google maps works? The core idea of google maps functionalities is based on machine learning techniques. Analyzing routes is important to transportation industry as it brings up advantages such as cost efficiency and increased profitability to different delivery companies and transportation organizations.

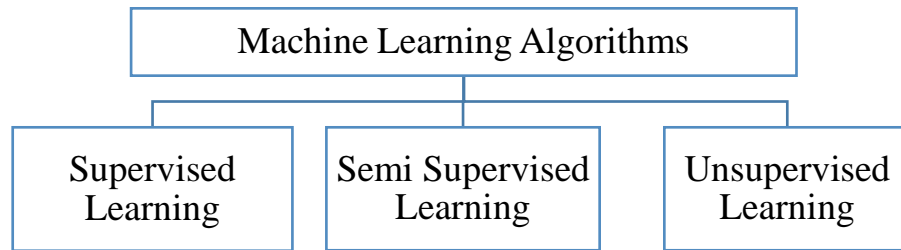


Figure 1 Machine Learning Algorithms

2.1.1 Supervised Machine Learning.

In supervised machine learning machines, machines are trained using labelled data, and based on that data, the outcome is predicted by the machine. Labelled data are data inputted that have already been tagged to the correct output. The concept of supervised machine learning corresponds to the concept of a student learning in the supervision of the teacher. It basically teaches the training data provided to the machines in order to predict the outcome correctly [7]. The objective of supervised learning algorithm is to find a mapping function to map the input variable x with the output variable y . Uses of supervised machine learning are found mostly on image classifications, fraud detections, spam filtering, risk assessment etc. There are many advantages when it comes to supervised machine learning as models are able to predict the output based on previous tests that have been done. In this kind of machine learning, when it comes to classes of objects, we have an exact idea on the matter. The reason we cannot use supervised machine learning in our model, is because SML is not suitable in this case. In this dissertation we are dealing with complex tasks, and supervised learning cannot predict the output as the data is different from the training set.

2.1.2 Semi - Supervised Machine Learning.

Just as you can see from the name, semi-supervised learning is split between unsupervised machine learning and supervised learning. In fact, most semi-supervised learning strategies are based on extending either unsupervised or supervised learning to include additional information typical of the other learning paradigm [8]. Semi-Supervised Learning is very practical. It can use both labeled and unlabeled data to have a better performance than supervised learning. Using Semi – Supervised learning can lead to cost efficiency as a model, as it can reduce the annotation effort by having less labeled instances. Applications of Semi Supervised learning contains different things such as classification of content on the internet, when classifying web pages in search engine, image and audio analysis, and a new application of classification of protein sequences. This as a model, offers a moderation between advantages and disadvantages of supervised and unsupervised learning. The usage of Semi – Supervised model is going to increase in the following years.

2.2 Unsupervised Machine Learning.

This paper presents an unsupervised learning algorithm by using association rules mining techniques to find correlations between different words in tweets. Unsupervised Learning are designed as a technique that it doesn't need to be supervised by the user. The model works on its own to discover patterns and information that has not been detected previously. All the variable used in the analysis of unsupervised machine learning can be used for clustering and association mining techniques [9]. Unsupervised Learning algorithms include detection of anomalies, neural networks, but the main and the ones that we need for this thesis, is clustering. Sometimes unsupervised learning methods are more unpredictable than the other learning methods. It mainly deals with unlabeled data. Ultimately, unsupervised clustering learning identifies inherent groupings inside the unlabeled data, and it gives label to each data value. Then, by using association mining algorithms, it tends to represent the relationships between attributes.

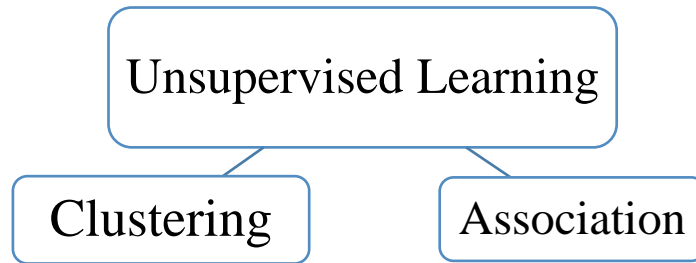


Figure 2 Types of Unsupervised Learning

2.2.1 Clustering

Clustering is a data mining technique that uses the similarities between unlabeled data to group them. This kind of algorithm are generally used to process raw, unclassified data objects into groups represented by structures or patterns in the information [10]. Some of the main clustering algorithms are as below:

2.2.1.1 K – Means Clustering

This is the most common way of Clustering methods. In this kind of method data points are distributed into data points called as K clusters. This process main aim is to find useful information from data. This is performed to model the time series behavior and to observe the pattern of the clustering. K-Means is mostly used in market segmentation, document clustering, image segmentation and compression.

2.2.1.2 Hierarchical clustering

Hierarchical Cluster analysis is divided into two methods, Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering. Divisive Clustering consists of partitioning one big cluster into smaller clusters. Agglomerative Clustering is a “bottom-up” approach, where observations starts in their own clusters, and pairs of

clusters are merged as one moves up the hierarchy. There are several ways of implementing Hierarchical clustering, some of them include Orange, R, Weka, SciPy (Python) etc.

2.2.2 Anomaly Detection

Just like it is called, anomaly detection is a method used to distinguish rare and unusual events. Every new data item that comes in, is observed in order to see if the normality level is normal, and if it is not, then an anomaly score is determined. In other words, if the level of the deviation is higher than the predefined threshold, then this data is considered to be an anomaly data. From there it is simpler to work with the data. This kind of method is mostly used in medicine, specifically in medical images, and it benefits doctors to closely check situations with patients.

2.3 Association Rules Mining

Association Rules Mining, or ARM is an unsupervised learning method. This kind of method allows you to do discovery. Usually, many researchers refer to unsupervised learning as a lacking prejudice method [11]. ARM has a lot to do with transaction data. Let's have a look on the classic review on how Association Rules work [12].

A transaction is a collection of items, and when you have a dataset of transactions, each row (if you have it in basket format, represents a set of items. Association rule mining is based on set theory. What you can do with this dataset is you can create what is called rules. Interestingly enough one of the most common rules is, if somebody goes to the supermarket and they put diapers in their chart, as it turns out, there is a very high probability that they are also going to put beer in their shopping chart.

2.3.1 Market Basket, an Association Rule Mining Practice

The scope of this thesis relies on the association rule mining. In order for us to understand deeply how association rule mining works, let's have a look at the classic example of association rule practice, which is called market basket. Market Basket Analysis is a practice which is widely used in retail, where transactions data from customer purchases are taken, and that data is used to find correlations, or associations between items in transactions. Let us use a dataset of 5 transactions to better understand the main idea of it [13]:

Table 1 Examples of Association Rules [13].

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

From those transactions in the dataset above, some rules are created as below:

$\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$

$\{\text{Milk, Bread}\} \rightarrow \{\text{Coke}\}$

$\{\text{Milk, Bread}\} \rightarrow \{\text{Coke, Diaper}\}$

$\{\text{Diapers}\} \rightarrow \{\text{Beer, Bread}\}$

This is what we start with when we talk about Association Rule mining, but the real question is, what things are associated and why do we care? The classic example leads to decision making when it comes to businesses as decision makers make plans based on information and based on data, and this is what we care about. Now taking this example a

step further, in order to answer questions that arise when it comes to decision making, measures are available in association rules in order to understand the data more. These measures of Set Correlation are mentioned as below. Sets means that any rule that can be generated is made up of a set of items, let's call it a set X (and it can be empty), and it goes to the set of Y (which can also be empty). The question that derives from here is, if we have X, what is the probability that we're going to have Y?

2.3.2 Common Measures in Association Rule Mining

There are three common measures that we use for Association Rule Mining:

Let X and Y be sets and assume rule $X \rightarrow Y$

$$Support = \frac{freq(X, Y)}{N}$$

Equation 1 Support

Support is just based on the probability of both of the sets occurring together divided by the total number of transactions. It is a good and helpful measure, but it is not enough because it can happen that one of those items is something that everybody buys all the time if you're dealing with items, then it's going to be with everything, and its measure is not going to be very accurate. Then what if you have items that come together all the time but are not very popular in general.

The other measure is called Confidence:

$$Confidence = \frac{freq(X, Y)}{freq(X)}$$

Equation 2 Confidence

Confidence is basically conditional probability. It is the probability that these items occur together given that the person has the set of items X in their cart. This is a better measurement than Support.

The last but not least measure is called Lift:

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Equation 3 Lift

Lift is the measure of dependent or correlated events. Association rules should have >1 lift to be meaningful. In lift equation, if the numerator and the denominator are the same, that's independence, meaning that these items in X and Y they don't really go together.

If the lift is between zero, it would be disjointed, and one, it would be independence, then the rule is not meaningful. This is really important, as rules can be generated which will have a high support, and think that they are meaningful, and then when realizing that the lift=1 that rule is "disqualified", because it is not meaningful [14].

These 3 measures are called interestingness [9] - rules that are interesting, and rules that are not interesting. These three different measures help determine which of these rules are maybe more or less interesting.

If we go back to Table 1, the example we see there is pretty much a common work [12] on the basic Market Basket analysis to help us understand more on the way the measures work and which ones to rely on, when it comes to importance and/or interestingness [9]. Table 1 has a total of five transactions. Since we've earlier mentioned the relation between beer and diaper, let's stick to that to have a closer look on the equations mentioned above. Beer and Diaper occur together twice, out of a transaction set of 5 transactions.

Given: {Diaper} → {Beer}

$$\text{Sup}(\{\text{Diaper}\}, \{\text{Beer}\}) = 2/5 = .40 = \mathbf{40\%}$$

This means that Diaper and Beer occur twice out of 5 transactions, so the Support is 40%

$$\text{Conf}(\{\text{Diaper}\}, \{\text{Beer}\}) = P(\{\text{Diaper}, \{\text{Beer}\}\}) / P\{\text{Diaper}\} = (2/5)$$

$$= P(\{\text{Diaper}\}, \{\text{Beer}\}) / P(\{\text{Diaper}\})$$

$$= (2/5) / (3/5) = \mathbf{66.7\%}$$

So in this case, the probability that these items occur together is 66.7% given that the person has diapers in their cart. This measurement gives us a better understanding on the probability that those items occur together.

Let's have a look at the case of measurement Lift:

$$\text{Lift}(\{\text{Diaper}\}, \{\text{Beer}\}) = \text{Sup}(\{\text{Diaper}\}, \{\text{Beer}\}) / \text{Sup}(\{\text{Beer}\}) * \text{Sup}(\{\text{Diaper}\}) =$$

$$(2/5) / (3/5) * (3/5) = \mathbf{1.11}$$

As you can see, Lift is greater than 1, and this means that the relation between those two items, in this case Beer and Diapers, are meaningful. If it had been equal to 1 or less than one, in that case those two items would have not been meaningful, or not at all interesting to be considered that they go together well.

2.4 The Apriori Algorithm

In R coding and in Python there are packages that allow us to perform what's called the Apriori algorithm. This kind of algorithm was introduced by Agrawal in 1994 and it is known to be as one of the best achievements in the history of mining association rules [15]. This kind of association rule is by far the most famous association rule algorithm [16]. What Apriori does is it selects candidate items by joining the big item sets of the previous pass and deleting the small ones from the previous pass. What happens here is that just because it selects large item sets from the previous pass, the number of candidate large item sets is significantly reduced [16].

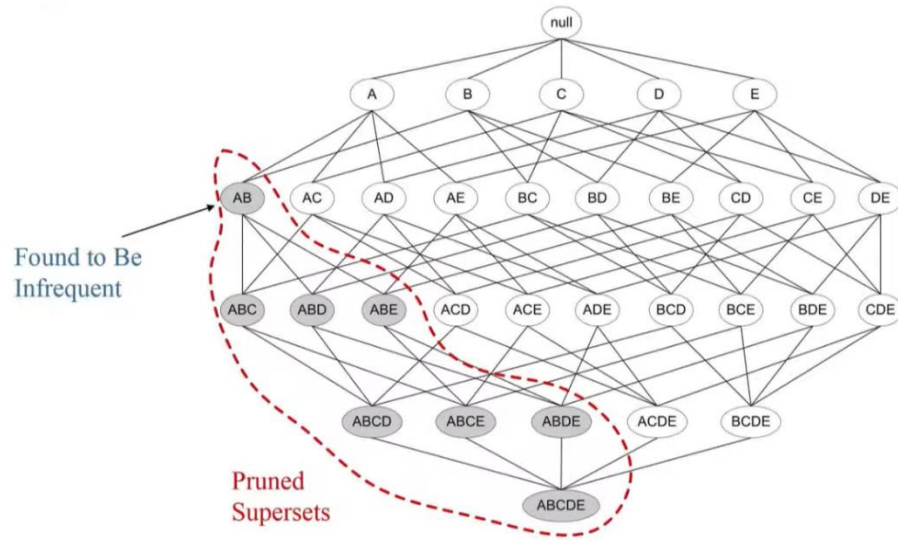


Figure 3 The Apriori Algorithm

What this basically does, you can imagine an enormous set of items, and you want to look at every possible combination of every item, how huge that tree is going to be. Figure 1 shows an example of that, and that example is a set of only 5 items. From those 5 items we observe a huge number of combinations. Apriori prunes based on supersets. So it looks at what is some kind of threshold of some measurement, say set A and set B, where A can be multiple items and B can be multiple individual items. If it turns out that they are infrequent, or they don't occur much at all, we prune everything that is a superset of that, because by definition, they will not occur a lot. And that is how R works the Apriori, because if we had to do manual work, we would have to look at the graph to see frequencies, and it would take a lot of time.

2.5 R Programming

For this study, R Studio version R 3.6.3 will be used. R Studio is a freeware for integrated development environment of R programming language [17]. There are two versions available, R Studio Desktop and R Studio Server. R programming is written in C++ and its packages are implemented in the R Studio. R programming is a machine learning programming language used extensively in the field of data mining. This programming language is an unsupervised non-linear algorithm to observe how item sets are associated with each other. Retailers use this kind of programming in grocery stores, supermarkets that have large databases of transactions. It is basically what social media like Facebook and Instagram do to track what you purchase by these recommendation engines.

The other side of the coin is: How do we make it work in R? Everything we have mentioned above has all been interesting theories. There are three general ways on how to represent the transaction data that we have seen so far:

Table 2 Reviewing Association Rules [13]

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

This type of format which we already saw earlier (which is called basket).

Table 3 Second way of data representation [13]

1	Bread
1	Coke
1	Milk
2	Beer
2	Bread
3	Beer
3	Coke
3	Diaper
3	Milk
4	Beer
4	Bread
4	Diaper
4	Milk
5	Coke
5	Diaper
5	Milk

The second way of data representation is called single.

Table 4 Third way of data representation [13]

TID	Bread	Coke	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

The third way of data representation is called sparse matrix. The format in table 3 is interesting because even when you store it in a CSV file it is not a record data, and it matters because we do not want it to be record data. As we have mentioned several times so far, association rules are an unsupervised learning method, meaning that we have to do with unlabeled data and users do not need to supervise the model. In our model, in order to get to the main data set, we have to do a bunch of different steps so that we can see proper results.

CHAPTER 3

METHODOLOGY

In this Chapter, a detailed description of the methodology of the thesis will be shown. The aim of this thesis is to analyze a twitter data set based using association rules and determine the relation words have between them.

The methodology of this thesis is based on “market basket” analysis, using association rules methods such as support, confidence and lift by using R programming

3.1 Introduction

Social Media is an everyday practice for millions of users all over the world on a daily basis. Interests that users talk about in social media is an important thing to observe, as many businesses are advertising their products through social media as their primary way of marketing. So, in this paper we are going to use twitter data to see the associations between things that people talk about. First, we look at the raw data and do some cleaning and formatting of it in order to be able to work on that data. Then we remove specific words that are not relevant or do not make any sense, as it affects our model results. Furthermore, we test the clean dataset to see the associations between tweets.

3.2 Data Set Used

The tweets extracted for this study have been fetched using Twitter Developer Platform to collect the data from tweets that contains words like chocolate. An amount of 100 chocolate related tweets have been taken into consideration for this study. In the upcoming sections, it will be explained why this small number of tweets have been

selected for the study. To have access to twitter data programmatically, we need to create an app on Twitter Developer that interacts with the Twitter API. After creating this app, the platform will provide us with consumer key and consumer secret key, which are classified confidential and should not be shared with others. Similarly, the strings should also be kept private: they proved the application access twitter on behalf of an account.

1.3 Specifications on the tweets data

Knowing the history of how association rules are used and where they are used, everyone uses those rules on transactions. It would be interesting to do the same for twitter data, and see if there are specific words that are related to each other. In order to have better results from our code, we have to keep in mind four main points:

- 1) We do not want “a bag of words” or a table of word frequencies.
- 2) We will need to create a “set of transactions” – one for each tweet.
- 3) Each row will contain a tweet.
- 4) Each column is a word (token) in that tweet.

Table 5 How tweets should look like.

How	we	intend	tweets	to	look	like
Computer	engineering	is	awesome			

Table 5 gives us a better visualization on how we want our tweets to look like in our .csv files where each row contains a sentence (tweet), and each column contains a word (token).

We don't want the tweets to be a bag of words. We want to make sure to retain each tweet, as its own transaction, just like the transactions in the classic market basket analysis.

In order to make that happen first we have to see which libraries we are going to use in R for association rule mining.

```
Library(arules)
Library(rtweet)
Library(twitter)
Library(ROAuth)
Library(jsonlite)
#library (streamR)
Library(rjson)
Library(tokenizers)
Library(tidyverse)
Library(plyr)
Library(dplyr)
Library(ggplot2)
#install.packages("syuzhet")
Library(arulesViz)
```

Listing 1

3.4 Data Collection

In order to take the dataset from twitter, firstly a Twitter Developer Account has been set up. This step is done by going to Twitter Developer Website. In order to do anything with Twitter you have to have access for a developer account.

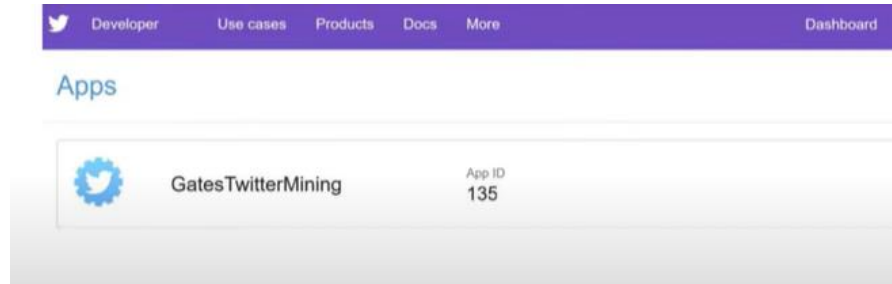


Figure 4 Developer Twitter Credentials

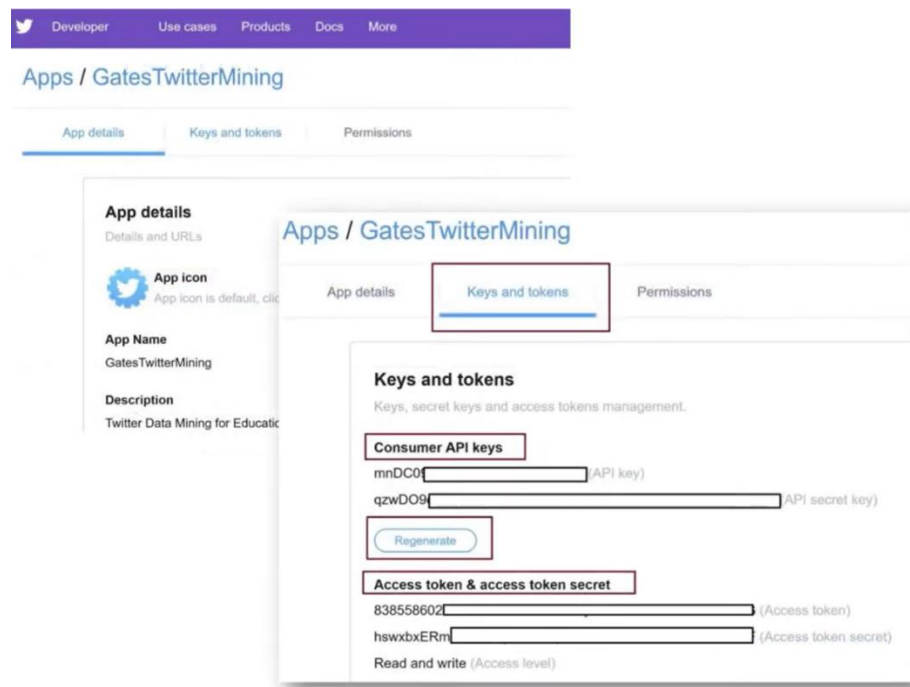
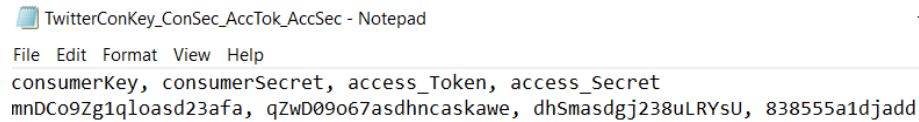


Figure 5 Getting the keys and tokens

The API works as a “middleman” between two pieces of software, as there are many functions that can be used in R programming in order to retrieve data from the app created in the API. All of the keys and tokens should be received and then be to put into R.

3.4.1 R Twitter Options



TwitterConKey_ConSec_AccTok_AccSec - Notepad
File Edit Format View Help
consumerKey, consumerSecret, access_Token, access_Secret
mnDCo9Zg1qloasd23afa, qZwD09o67asdhncaskawe, dhSmasdgj238uLRYsU, 838555a1djadd

Figure 6 Twitter Keys and tokens in .txt

A text file has been created containing the variable names, and under them the actual values of the tokens extracted from Twitter Developer. In this way the data can be easily readable and easily updatable to be used in multiple files, so instead of writing a big code every time we want to run it, it lives in that file.

```
Filename="TwitterConKey_ConSec_AccTok_AccSec.txt"  
(tokens<-read.csv(filename, header= TRUE, sep= “,”))  
(consumerKey=as.character(tokens$consumerKey))  
consumerSecret=as.character(tokens$consumerSecret)  
access_Token=as.character(tokens$access_Token)  
access_Secret=as.character (tokens$access_Secret)
```

Listing 2

Based on the extracted values and tokens from the Twitter Developer above, we are going to put to use the library twittR to pull that into twitter.

```
# Using twittR#  
setup_twitter_oauth(consumerKey, consumerSecret, access_Token, access_Secret)  
Search<-twittR::searchTwitter("#IloveChocolate",n=100,since="2018-09-09")  
(Search_DF <- twListToDF(Search))  
TransactionTweetsFile= "Choc.csv"
```

Listing 3

This code wraps the OAuth authentication from the httr package for a twitter session. All of the arguments are supplied by Twitter, and with the access tokens previously mentioned above, we will be able to generate results from the function searchTwitter. In the above code you can see that we have put a condition to search the supplied string “I love Chocolate”, from which we intend to find rules and correlations with other words. The maximum number of tweets to return has been given a value of 100, and to return tweets not older than September 9th 2018. The search_DF function will take the list of objects from the twittR class and return a data frame version. And lastly, we will save all the returned tweets to a .csv file which we will call “Choc.csv”

Let’s take a look at the returned list of objects which are saved at the “Choc.csv” file.

(Search_DF <- twListToDF(Search))

```

1 The other day I woke up craving chocolate cupcakes. Today I'm craving @HersheyCompany chocolate bars.
  think the u... https://t.co/NtGH4eaSRC
2 WHO SAID "CHOCOLATE"? \n
  ovechocolate... https://t.co/DzzmvJlKEh
3 @ClaireValy @LowngSnake @Firebox #ILOVECHOC
  ATE\nI love Chocolate very very much.
4 #HealthTips #momlife #sahmlife #toddlers #ilovechocolate #homeschoolmom #bethechange
  oingitformygirls #fitmom #feeltheburn
5 RT @Kelly_Hawrylysh: #Fairtrade sourcing needed more than ever to avoid chocapocalypse!!! https://t
  o/dbxw3eQfTc #SDG12 @FairtradeAfrica...
6 RT @Kelly_Hawrylysh: #Fairtrade sourcing needed more than ever to avoid chocapocalypse!!! https://t
  o/dbxw3eQfTc #SDG12 @FairtradeAfrica...
  FAVORITED FavoriteCount replyToSN created truncated replyToSID
1 FALSE 0 <NA> 2018-09-27 12:12:52 TRUE <NA>
2 FALSE 0 <NA> 2018-09-27 10:51:42 TRUE <NA>
3 FALSE 0 ClaireValy 2018-09-27 00:45:43 FALSE 1044897146326208513
4 FALSE 0 templin_katie 2018-09-26 19:49:55 FALSE 1045037612388536321
5 FALSE 0 <NA> 2018-09-26 16:24:22 FALSE <NA>
6 FALSE 0 <NA> 2018-09-26 16:23:42 FALSE <NA>
  id replyToUID
1 1045285140505735169 <NA>
2 1045264712118734848 <NA>
3 1045112213915226113 2878148959
4 1045037771050618881 1035584652722036736
5 1044986045975220224 <NA>
6 1044985877456392194 <NA>
  statusSource
1 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
2 <a href="http://instagram.com" rel="nofollow">Instagram</a>
3 <a href="http://twitter.com" rel="nofollow">Twitter web client</a>
4 <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
5 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
6 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
  screenName retweetCount isRetweet retweeted longitude latitude
1 RachelTBue 0 FALSE FALSE <NA> <NA>
2 Niklaus_R 0 FALSE FALSE 4.35008 50.845
3 saminaseem16 0 FALSE FALSE <NA> <NA>

```

Figure 7 Returned list of objects.

When you look at the objects, if you are familiar with twitter at all it is going to look very familiar. The reason why we ran these objects in data frame version is because we only want to access the text of these tweets. The key here is that each tweet must be a transaction and each word (token) in the tweet should be in its own column.

To have a clean dataset, a file has been created, and the very first tweet as a transaction has been put inside that file and then separated as .csv:

```
> (Search_DF$text[1])
```

```
[1] "The other day I woke up craving chocolate cupcakes. Today I'm craving @HersheyCompany chocolate bars. I think the universe wants me to eat more chocolate! #LoveChocolate"
```

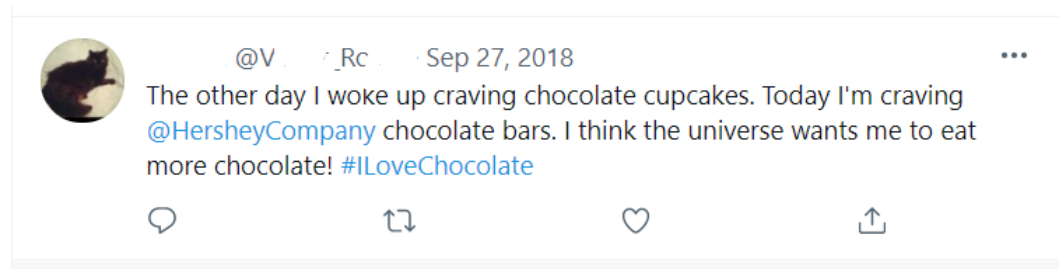


Figure 8 An example of how a tweet looks like.

3.5 Building the Transaction File

Tokens are used to break down human readable information into machine learning readable information.

```
## Start the file
Trans <-file(TransactionTweetsFile)

##Tokenize to words

Tokens<-tokenizers::tokenize_words(Search_DF$text[1],
stopwords=stopwords::stopwords("en"),
lowercase= TRUE, strip_punct = TRUE, strip_numeric= TRUE, simplify = TRUE)

##Write squished tokens

Cat (unlist(str_squish(Tokens)), "\n", file=Trans, sep=";")

Close(Trans)

##Append remaining lists of tokens into file

##Recall – a list of tokens is the set of words from a Tweet

Trans<- file (TransactionTweetsFile, "open = "a")

for (i in 2:nrow(Search_DF)) {

Tokens<-tokenize_words
(Search_DF$textpi[,stopwords=stopwords::stopwords("en").

Lowercase = TRUE, strip_punct= TRUE, simplify=TRUE)

Cat (unlist(str_squish(Tokens)), "\n", file=Trans, sep= ";")

}

Close (Trans)
```

Listing 4

In the code above, tokens have been used to build the transactions file. Tokenization breaks down human readable text into machine learning readable components. Then we squish the values into range so that when we run the code and open excel file where the changes were made to, we observe a messy file. The main point in this piece of code is that all tweets are separated into words.

day	woke	craving	chocolate	cupcakes	today	craving	hershey	chocolate	bars	think	u	https	t.co	ntgh4e	
said	chocolate		feed	feedsmart	honey	welovechocola	https	t.co	dzzmvykeh						
ckaireval	lowngsnak	firebox	ilovechocola	love	chocolate	much									
healthtip	momlife	sahmlife	toddlers	ilovechocc	homeschool	bethechange	doingit	fitmom	feeltheburn						
rt	kelly_haw	fairtrade	sourcing	needed	ever	avoid	chocapocaly	https	t.co	dbxw3	sdg12	fairtradeafrica			
rt	kelly_haw	fairtrade	sourcing	needed	ever	avoid	chocapocaly	https	t.co	dbxw3	sdg12	fairtradeafrica			
cada	dia	estamos	mas	listos	para	navidad	taza		3	pack	de	venta	en	cityclu	navidad
fairtrade	sourcing	needed	ever	avoid	chocapocalyp	https	t.co	dbxw3eq	sdg12	https	t.co	rgmtaombom			
ilovecho	chocolate	adictaalch	https	t.co	kpzofu8ix2										
see	big	chocolate	show	Saturday	night	ilovechocolate									
else	can	say	thehouse	braziliantruff	truffles	brigadeiro	desserts	https	t.co	pzayia63ir					
touch	cocoa	please	ilovechocalt	bless	https	t.co	vx7v7csfr5								
bako_nw	weekend	choc	dome	hiding	double	chocolate	cheesecake	ilovechoco	https	t.co	f2ginuvfq				
ilovechocolate															
los	lunes	lucen	tan	malos	si	los	ves	con	la	actitud	corre	chocola	iniciod	felizlun	
enough	words	express	thankful	amazing	coworkers	thank	thank	onl	https	t.co	2glightudgg				
casa	ino	nostra	przedstaw	hotel	hotelwgor	tatry	zakopane	nowy tang	deser	slody	suflet	https	t.co		
rt	ccfchocola	crunchy	biscuit	dipped	chocolate	foodporn	foodporn	sweets	instafood	instafo	food	selicious	choco	deser	
crunchy	biscuit	dipped	chocolate	foodporn	yummy	sweets	sweets	instafood	delicious	delicio	choccc	dessert	https	t.co	
bbciami	light	ilovechocolate													

Figure 9 Tweets into tokens opened with Excel

Like we've mentioned earlier in this chapter, the main focus of our work was to be able to turn each tweet into "transactions", like the transactions used in "market basket" analysis in order to proceed with the results. Table 6 shows the first 20 rows of the tweets, where every row is a tweet, and every column is a word, or a token.

If you look closely to the table, you will notice that there are many tokens that do not make any logical sense. As many of us already know, tweets are usually made up of sentences, and also users tend to add attachments, such as pictures, links etc. This kind of information is not valuable for our model because it will mess up the rules. In table 6 we observe many tokens that come from links of the tweets of twitter users.

3.6 Data Cleaning

In order to have a better and cleaner dataset, we will have to read and inspect the “transactions” and then summarize it, using R code.

```
### Read in the tweet transactions

TweetTrans <- read.transactions (TransactionTweetsFile, rm.duplicates=
FALSE, format = “basket”, sep = “,”)
```

Listing 5

This code reads the tweet transactions. The function `rm.duplicates` is set to false, as that function is asking to remove duplicates, and we do not want that, as we need to see the frequencies of the transactions in the data set. We also want the format to be in basket, and each line of the file to be separated by a comma.

```
[59] {1,
    along,
    box,
    chocolates,
    days,
    domme,
    findom,
    finsub,
    godiva,
    ilovechocolate,
    pay,
    send}
[60] {chocolate,
    delicious,
    food,
    foodporn,
    https,
    instafood,
    introducing,
    love,
    mango,
    marzipan,
    sweets,
    t.co,
    truffles,
    u17wpqhxx,
    yummy}
```

Figure 10 `read.transactions` in basket format

```

Inspect(TweetTrans)
##See the words that occur the most
Sample_Trans <-sample(TweetTrans,50)
Summary (Sample_Trans)

```

Listing 6

This function opens up a window that offers several views to analyze the series. If changes are done, adjustments will be calculated. In this code we analyze the summary of the 50 words that occur the most.

Table 6 Most Frequent Items.

```

most
frequent items:
      https  t.co  chocolate  ilovechocolate  rt
      35     35     25           23           9

```

As we discussed earlier, Figure 7 shows us the most frequent items from the data set, and as we notice, there are a lot of tokens which we do not need, such as https, which are the links extracted from the tweets, rt-s which are the Retweets from those tweets etc. We are going to have to take these away, so it doesn't destroy our association rules.

3.6.1 Actual Clean Up

Now that we know what the frequent items are, and we have it visually presented into basket format and separated by comma, we can proceed with the cleanup phase of the data set.

We want to read the transactions data from table 6 into a DF.

```
TweetDF <- read.csv(TransactionTweetsFile, header = FALSE, sep="",")
```


Head(TweetDF)

```
> TweetDF <- read.csv(TransactionTweetsFile, header = FALSE, sep = ",")
> head(TweetDF)
  v1      v2      v3      v4      v5
1  day    woke    craving chocolate cupcakes
2  said   chocolate feed feedsmartfood
3  clairevaly lowngsnake firebox ilovechocolate love
4  healthtips momlife sahmlife toddlers ilovechocolate
5  rt kelly_hawrylysh fairtrade sourcing needed
6  rt kelly_hawrylysh fairtrade sourcing needed
  v6      v7      v8      v9      v10      v11      v12      v13
1  today    craving hersheycompany chocolate bars think u https
2  honey welovechocolate https t.co dzzmvjlkeh
3  chocolate much
4  homeschoolmom bethchange doingitformygirls fitmom feeltheburn
5  ever avoid chocapocalypse https t.co dbxw3eqftc sdg12 fairtradeafrica
6  ever avoid chocapocalypse https t.co dbxw3eqftc sdg12 fairtradeafrica
```

Figure 11 Pre-Formatted transactions data into Data Frame

In order to remove words or tokens that we do not need, a combination of getting rid of stuff by hand, and taking away things like digits by using `grepl`. A logical list is created and applied to the data and deleting things that fall under the logical true or false.

3.6.2 Removing Specific Words

At first, we are going to convert all columns to char. This is done with the following command.

```
TweetDF<-TweetDF %>%
Mutate_all(as.character)
(str(TweetDF))
```

Listing 7

Then, removal of certain words is going to be applied as shown below:

```
TweetDF[TweetDF == "t.co"] <-""
TweetDF[TweetDF == "rt"] <-""
TweetDF[TweetDF == "http"] <-""
TweetDF[TweetDF == "https"] <-""
```

Listing 8

3.6.3 Cleaning with grepl

```
MyDF<-NULL
for(i in 1 : ncol(TweetDF)) {
  MyList=c() #each list is a column of logicals...
  MyList=c(MyList,grepl("[[:digit:]]", TweetDF[[i]]))
  MyDF<-cbind (MyDF, MyList) ##create a logical DF
  ## TRUE is when a cell has a word that contains digits
}
## For all TRUE, replace with blank
TweetDF[MyDF] <-""
(TweetDF)
MyDF<-NULL
MyDF2<-NULL
MyDF3<-NULL
for(i in 1 : ncol(TweetDF)) {
  MyList=c() #each list is a column of logicals...
  MyList=c(MyList,grepl("[[:digit:]]", TweetDF[[i]]))
  MyList2=c() ## for small words
  MyList2=c(MyList2,grepl("[A-z]{4,}", TweetDF[[i]]))
  MyList3=c() ## for large words
  MyList3=c(MyList3,grepl("[A-z]{12,}", TweetDF[[i]]))
  MyDF<-cbind (MyDF, MyList)
  MyDF2<-cbind (MyDF2, MyList2)
  MyDF3<-cbind (MyDF3, MyList3)
}
## For all TRUE, replace with blank
TweetDF[MyDF] <- ""
TweetDF[MyDF2] <- ""
TweetDF[MyDF3] <- ""
(head (TweetDF,10))
```

Listing 9

The grepl function returns a logical vector indicating whether a match was found or not. We want to take all of the possible tokens that make no logical sense in a sentence by grepl. By creating logical lists as above, we take away things like digits etc. For small tokens that matches exactly 4 characters from A to Z, we create a list, and then replace those words with blank. For large tokens that matches more than 12 characters from A to Z, we also create a list for those tokens and replace those words with blank values.

After cleaning with grepl, our transactions will look more organized, cleaner, and we observe to have more relevant data in our data set.

	v1	v2	v3	v4	v5	v6	v7	v8	v9				
1	looking	healthy	food	swaps				eathealthy	nutrition				
2			dates	stuffed	roasted	almonds	dipped	chocolate	weekend				
3	dates	stuffed	roasted	almonds	dipped	chocolate	weekend	indulgences	giftbox				
4		reese's	cups	mini's	they're	half	calories	regular	ones				
5		cuando		aburre		helado		vainilla	tenemos				
6	bigass	chocolate	happydiwali	corporate	tings	bigting			africa				
7		birthday	cake	chocolate	orange	shared	friends						
8	doubt		chocolate	person	tries	make	move	chocolate					
9	another	great	gift	idea	just	love	bottles	come					
10													
	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21	v22
1													
2	indulgences	giftbox	foodporn	yummy	sweets								
3	foodporn	yummy											
4	unless			they're									
5	siempre		chocolate	pero	blanco	pata	hacer	nuestro					
6	umhlanga												
7	birthday			gmakffbpaj									
8													
9		exciting	christmas										
10													

Figure 12 Formatted Transactions Data Frame

We can look at those results of the cleaner tweets as individual transactions in basket format.

```
    sena,  
[70] {chocolate,  
      delicious,  
      food,  
      foodporn,  
      instafood,  
      introducing,  
      love,  
      mango,  
      marzipan,  
      sweets,  
      truffles,  
      yummy}  
[71] {bali's,  
      big,  
      check,  
      chocolatiers,  
      ilovechocolate,  
      six,  
      theyakmag,  
      theyakmagazine,  
      yak}
```

Figure 13 Tweets as transactions, basket format

CHAPTER 4

RESULTS AND DISCUSSIONS

In this chapter we will discuss on the results based on the methodology used in the previous chapter. So far, we have been able to extract the data from Twitter Developer, used R programming to read the tweets and convert them into tokens. Then some cleaning of the data has been done to remove words that are unnecessary and mess up the association rules. With a clean data set, it is now possible to implement the association rules in R and see correlation between words.

4.1 Implementing Association Rules in R

Finally, we are able to look at the Association Rules.

```
TweetTrans_rules = arules::apriori(TweetTrans, parameter = list(support=.0001,
confidence=.0001, minlen=2, maxlen=6))
inspect(TweetTrans_rules([1:30])
##Sorted
SortedRules_sup <- sort(TweetTrans_rules, by="support", decreasing=TRUE)
Inspect(SortedRules_sup[1:20])
```

Listing 10

In this code, we look at the association rules from the dataset, and we mine rules with minimum support of 0.0001, and minimum confidence of 0.0001. The “maxlen” function defines the maximum number of items in each itemset of frequent items. Apriori only creates rules with one item in the RHS [15]. The default value for minlen is 1. This means that rules with only one item (i.e., an empty LHS) like :

{ } => {chocolate}

will be created. These rules mean that no matter what other items are involved, the item in the RHS will appear with the probability given by the rule's confidence which equals the support. So, in this case we want to avoid these rules so we used the argument (minlen=2).

The results we got from the rules are what we expected them to be.

```
> inspect(SortedRules_sup[1:20])
  lhs          rhs      support  confidence lift    count
[1] {national} => {chocolate} 0.11267606 1.00000000 1.731707 8
[2] {chocolate} => {national} 0.11267606 0.19512195 1.731707 8
[3] {dessert} => {chocolate} 0.07042254 1.00000000 1.731707 5
[4] {chocolate} => {dessert} 0.07042254 0.12195122 1.731707 5
[5] {foodporn} => {chocolate} 0.07042254 1.00000000 1.731707 5
[6] {chocolate} => {foodporn} 0.07042254 0.12195122 1.731707 5
[7] {happy} => {national} 0.05633803 0.66666667 5.916667 4
[8] {national} => {happy} 0.05633803 0.50000000 5.916667 4
[9] {happy} => {chocolate} 0.05633803 0.66666667 1.154472 4
[10] {chocolate} => {happy} 0.05633803 0.09756098 1.154472 4
[11] {weekend} => {chocolate} 0.05633803 1.00000000 1.731707 4
[12] {chocolate} => {weekend} 0.05633803 0.09756098 1.731707 4
[13] {sweets} => {chocolate} 0.05633803 1.00000000 1.731707 4
[14] {chocolate} => {sweets} 0.05633803 0.09756098 1.731707 4
[15] {giftbox} => {yummy} 0.05633803 1.00000000 17.750000 4
[16] {yummy} => {giftbox} 0.05633803 1.00000000 17.750000 4
[17] {giftbox} => {foodporn} 0.05633803 1.00000000 14.200000 4
[18] {foodporn} => {giftbox} 0.05633803 0.80000000 14.200000 4
[19] {giftbox} => {chocolate} 0.05633803 1.00000000 1.731707 4
[20] {chocolate} => {giftbox} 0.05633803 0.09756098 1.731707 4
```

Figure 14 Sorted Rules

Let's have a close look at the Lift of Figure 14. The rules that have a high Lift are the rules that we should pay attention to most, as they are the rules which are most meaningful. We've gotten things like yummy, gift box and it these rules are also affected by the time it is done, whether it is Halloween, Christmas or any other holiday of the seasons. Also, another interesting rule that we observe from the results above is that there is a community that associates chocolate with foodporn.

4.2 A quick Plot

```
Library(arulesViz)
SortedRules_sup <- sort(TweetTrans_rules, by="support", decreasing=TRUE)
Inspect (Sorted Rules_sup[1:20])
Plot (SortedRules_sup[1:25], method="graph",
      engine='interactive', shading='confidence')
```

Listing 11

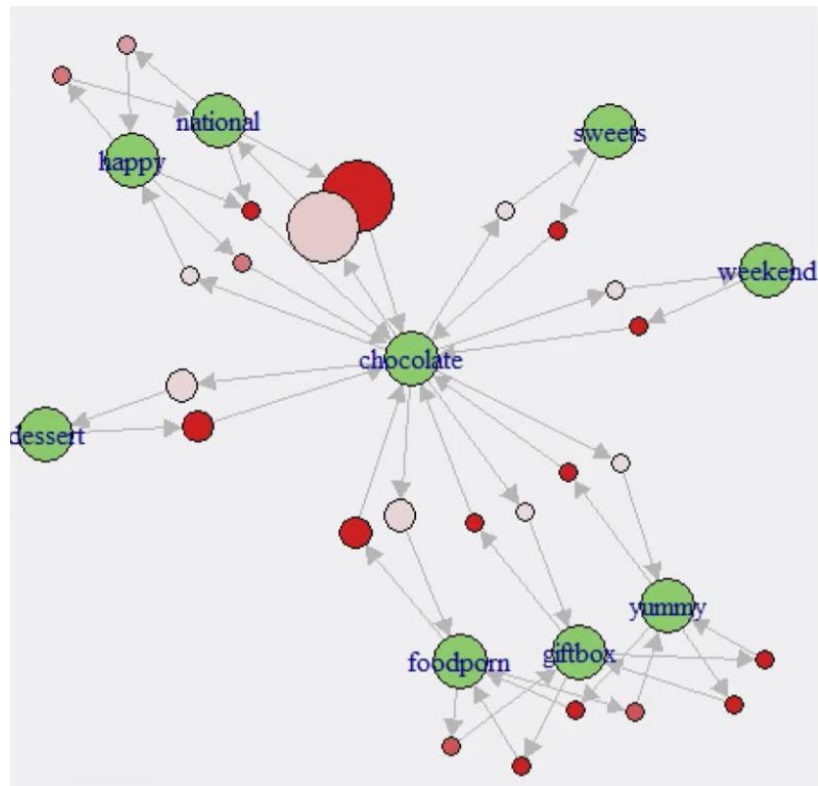


Figure 21 ArulezViz, a Quick Plot

In this arulezViz graphic we have a better view of which items are associated to each other. The arulezViz function creates the basic plot and it gives you a visualization. The size of the items in this graph is based on support, while the color is based on confidence, where darker is the higher confidence. Based on the results it is discovered that there is a very strong relation between the rules created. Chocolate is happy, it is a dessert, it happens on the weekends, it is associated with food pornography and sometimes it comes in gift boxes.

CHAPTER 5

Conclusions and Future Work

5.1 Conclusions

In this study we pose a few questions which are related to association rule mining. How can raw and unstructured data such as tweets be converted into meaningful data to be processed? What machine learning techniques can be efficient when finding associations between words in a set of Tweets? Association Rules have shown to be a success when it comes to transactional data, as using confidence, support and lift in the classic market basket analysis. In this research we tried to do the same market basket analysis, and this time not with “retail transactions”, but with tweets extracted from Twitter API.

In the preliminary analysis we demonstrated that the raw tweets can be tokenized into separated words and squished into a .csv file to have them transactional style. Then more data cleaning has been done to this database, as the model is not ready to be published yet. Manual work has been done to further clean the data set, as there were many words that needed to be removed, because there were words that had no meaning to the context of chocolate tweets. After cleaning with grepl, a much cleaner and organized dataset was formed, having more relevant information when playing with the dataset. At this point we have achieved the market basket style of data, knowing that market basket analysis consists of transactions. With R we have managed to convert full tweets into relevant words, and so converting them into “transactions”. Then association rules methods have been implemented. From the association methods used, we observe that the word chocolate is highly associated with the words happy, dessert, gift boxes and weekends.

5.2 Limitations and Future Work

The results to this thesis are limited by several reasons. The most important reason why this thesis is limited, is because of the small number of tweets sample selected from the Twitter API. The study consisted of a lot of manual work of cleaning the dataset, to have a formatted file with the relevant words. If a larger sample had been chosen, more manual work would have been needed to clean the dataset from unnecessary words to keep only the relevant ones. Even though results turned out to be meaningful and words turned out to be related together, the study can be called as biased, as the word “chocolate” is often described as of a positive sentiment. In order to have a more time efficient research and reduce the manual work and since there are so many, I would suggest using other algorithms for a better performance. For future work, it would be interesting to see larger samples of data with more relevant words and compare results.

REFERENCES

- [1] R. Ramli, A. M. N. Shahrul and M. Yusof, "Ontological-based model for human resource decision support system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 50-62, 2010.
- [2] S. R. Department, "Number of social network users worldwide from 2017 to 2025," 28 January 2021. [Online]. Available: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>.
- [3] K. Smith, "BrandWatch," 2 January 2020. [Online]. Available: <https://www.brandwatch.com/blog/twitter-stats-and-statistics/>.
- [4] J. E. Bradley, P. Korfiatis, Z. Akkus and T. K. Kline, "Machine learning for medical imaging," *RadioGraphics*, p. 37, 2017.
- [5] v. d. M. Rob, "Gartner," Gartner Inc, [Online]. Available: <https://www.gartner.com/newsroom/id/3598917>. [Accessed 2021 July 10].
- [6] H. Li, "Sas," [Online]. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html. [Accessed 7 7 2021].
- [7] S. Jaiswal, "Javatpoint," [Online]. Available: <https://www.javatpoint.com/supervised-machine-learning>. [Accessed 7 July 2021].
- [8] D. T. Brachman Ronald, "Introduction to Semi – Supervised Learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009, p. 23.
- [9] M. Hahsler, C. Buchta, B. Gruen and K. Hornik, "arules: Mining Association Rules and Frequent Itemsets," pp. 6-8, 2021.

- [10] Education and I. Cloud, "IBM," [Online]. Available: <https://www.ibm.com/cloud/learn/unsupervised-learning>. [Accessed 12 July 2021].
- [11] G. Amy, "Association Rules Mining with Tweets," 26 3 2019. [Online]. Available: <https://youtu.be/eOOhn9CX2qU>. [Accessed 20 6 2021].
- [12] S. M. Pang-Ning Tan and A. Karpatne, *Introduction to Data Mining*, 2018.
- [13] Y.-L. Chen and e. al, ""Market basket analysis in a multiple store environment.", "*Decision support systems*, pp. 339-354, 2005.
- [14] A. Rakesh, T. Imieliński and A. N. Swami, "Mining Association Rules Between Sets of Items in Large Databases, . in International Conference on Management of Data," in *ACM Press*, 1993.
- [15] P. Badri, N. Gupta, R. Karn and Y. Rana, "Optimization of Association Rules Mining Apriori Algorithm Based on ACO," *International journal on emerging technologies*, vol. II, pp. 87-92, 2011.
- [16] M. Dunham, Y. Gruenwald and Z. Hossain, "A SURVEY OF ASSOCIATION RULES," 2001.
- [17] H. Shakir, A. Ribal, K. Amirrudin and H. Jiten, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study.," *Silhavy R. (eds) Cybernetics and Algorithms in Intelligent Systems. CSOC2018 2018. Advances in Intelligent Systems and Computing*, vol. 765, 2019.
- [18] F. Namugera, R. Wesonga and P. Jehopio, "Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda.," *Comput Soc Netw*, p. 3, 2019.