**qwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnm**

|  |
| --- |
| Data Mining  Data Solution Institute  11/15/2018  K.P.I SheneshPerera: IDM |

**BLANK PAGE**

Contents

[Introduction to Data Mining 3](#_Toc3489365)

[Historical background of data mining 5](#_Toc3489366)

[The process of data mining in-depth 5](#_Toc3489367)

[1. Data Selection 5](#_Toc3489368)

[2. Data Transformation 6](#_Toc3489369)

[3. Data Mining 6](#_Toc3489370)

[4. Final result evaluation 6](#_Toc3489371)

[Traditional data mining techniques & approaches 6](#_Toc3489372)

[2. Regression (statistical based) 7](#_Toc3489373)

[Modern data mining techniques & approaches 7](#_Toc3489374)

[1. Classification 7](#_Toc3489375)

[2. Decision Trees 7](#_Toc3489376)

[3. Neural Networks 8](#_Toc3489377)

[4. Association rule 8](#_Toc3489378)

[5. Clustering 8](#_Toc3489379)

[Popular data mining tools 8](#_Toc3489380)

[ Rapid Miner 8](#_Toc3489381)

[ Oracle Data Mining 8](#_Toc3489382)

[ Orange 9](#_Toc3489383)

[ Weka 9](#_Toc3489384)

[The building blocks of most traditional & modern approaches/techniques 9](#_Toc3489385)

[Scopes of data mining 10](#_Toc3489386)

[Data Mining algorithms & how they work 10](#_Toc3489387)

[k-means algorithm implementation 11](#_Toc3489388)

[Text mining & how it works 15](#_Toc3489389)

[Text mining techniques, methods and approaches 15](#_Toc3489390)

[Text mining algorithms and how they work 15](#_Toc3489391)

[Real world examples of text mining 16](#_Toc3489392)

[Real world implementation 16](#_Toc3489393)

[Code & Output 17](#_Toc3489394)

[Graph mining 17](#_Toc3489395)

[Graph mining algorithms 18](#_Toc3489396)

[Real world examples of Graph Mining 18](#_Toc3489397)

[Real world implementation 18](#_Toc3489398)

[Code 20](#_Toc3489399)

[Ouput 20](#_Toc3489400)

# Introduction to Data Mining

This is an introduction to data mining, its scope, purpose and how it works. Data mining could be defined as “The process of obtaining hidden predictive information”.

Let’s start with the big picture first. This all starts with something called "Knowledge Discovery in Database". Data mining is basically one of the steps in the process of knowledge discovery in a database (KDD). Knowledge discovery process employs 5 main steps:

1. Selection
2. Pre-processing
3. Transformation
4. Data Mining
5. Evaluation

Selection is the phase where we identify the data, pre-processing is the phase where we cleanse and organize the data, transformation phase is required for preparation of data, and then you reach the data mining phase. Lastly we analyze information which is the evaluation phase to test the results of the data mining.

So why do we call this result knowledge? Why not information or data? This is because there are a lot of differences between the terms; data, information and knowledge itself.

The following is an example to explain what data mining is in a broad perspective;

You run a local departmental store and you log all the details of your customers in the store database. You know the names of your customers and what items they buy each day. John, Peter and Dave visit your shop every Monday and buy a lot of bread. You store this information in your database. This is known as data. Any time you want to which customers buy bread, you can query your database and get them, what you get is called information. You want to know how many loaves of bread are sold on each day in a particular month, you can again query your database appropriately and you’d get them that’s also information.

But suppose there are 1000 other customers who also buys lots of loaves of bread from your shop every Monday and all of them are vegetarian. Then using this information you can come to a conlusion that John, Peter and Dave are also vegetarian.

True, whether or not John, Peter and Dave are vegetarian is not known to as you have not obtained them as data. Therefore something like this cannot be retrieved from the database as information *but* you learnt and came to a conclusion indirectly, this is knowledge and is the very essence of data mining. This **entire** process of coming to a point where you can discover knowledge is collectively contributes to data mining.

# data_mining_timelineHistorical background of data mining

Data mining became a concept of wonder in the 1990s but the origins of this concept go way back to the 1700s. Data mining has its roots that go back its siblings which are classical statistics, A.I. and machine learning.

It initially started with Thomas Bayes’ theorem in the 1700s of identifying patterns in data then further on backed by techniques such as regression analysis in 1800s. The sheer demand, increasing efficiency and the popularity of computer technology now has increased the need for data collection, storage and much more importantly manipulations.

Figure 1.0, hackerbits.com, Ray Li, 12/05/2015

As data stores have grown both in size and complexity, direct approaches for data analysis became increasingly augmented with indirect and automatic data processing, however much later on this was improved by discoveries such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and vector machines (1990s).

Finally, one of the evolved form of data mining which deep learning is being explored in the present. It is said to be capble of capturing dependencies and extremely complex patterns way beyond other techniques, therefore challenging the biggest problems in the world of data mining, data science and A.I.

# The process of data mining in-depth

The previously described 5 steps in the process of data mining can be technically combined to 4 basic steps that must be present in any data mining procedure, namely the 4 basic steps are:

## Data Selection

A data warehouse is a large collection of a variety of data, not all of which is needed to achieve each the basic goals of data -mining. The first step in the data-mining process is to select the target data.

For example, making databases contain data describing customer purchases, demographics and life style preferences. To identify which items and quantities to purchase for a particular store, as well as how to organize the items on the store's shelves a marketing executive might need only to combine customer purchase data

with demographic data.

The selected data types may be organizing along multiple tables, during data selection; the user might need to perform table joins. Furthermore, even after selecting the desired database tables, mining the contents of the entire table is not always necessary for identifying useful information.

Under certain conditions and for certain types of data- mining operations for example when creating a classification of regression model, it is usually a less expensive (in terms of time taken and resources consumed) operation to sample the appropriate table, which might have been created by joining other tables, and then mine only the sample.

## Data Transformation

After successful data selection has been performed, the next task is to transform the data. This step cannot be performed without being aware of 3 very important contexts. The task to be performed, the data mining operations that will be performed and the data mining technique that will be used.

Transformation methods include organizing data in desired ways, and converting one type of data to another.

## Data Mining

The next step is to subsequently mine the transformed data using one or more techniquesto extract the desired type of information.

For example, to develop an accurate, “classification” model that performs prediction as to whether or not a subscriber to a particular service will renew their subscription(s) a company manager might need to employ a clustering technique to segment the subscriber database, then apply rule induction to automatically create a classification model for each desired cluster.

## Final result evaluation

Finally, the last step in the process is to verify if the result that has been obtained matches the desired goals and objectives that were set in order to be achieved through data mining. Then in order to prepare this result to an actual management level comprehensible manner, one must decide how best to visualize the results that have been obtained.

# Traditional data mining techniques & approaches

Traditional data mining techniques/approaches are those that are identified as the foundations of data mining, these techniques are very popular but usually not used in the industry as much as the modern ones nowadays but most modern techniques are modified versions of these techniques.

1. Naïve Bayes

Naive Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the “Maximum Posteriori” decision rule in a Bayesian setting.

It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

## 2. Regression (statistical based)

Regression is a technique by which you model the relationship between 2 variables. This technique is usually performed to identify the mathematical relationship between said variables, to see if they are related in a linear or non-linear manner.

Generally, if the relationship is linear we mean to say that the variables can be modelled using the equation of a straight line through an intercept in the Cartesian plane, which is y=mx+c, where the variable denoting y will be dependent and x independent.

Regression has its roots in a mathematical field known as statistics, where dynamic relationships are quantified and qualitied.

# Modern data mining techniques & approaches

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning.

In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Neural Networks, Association Rule Mining etc. are common examples of supervised learning.

In unsupervised learning, we do not create a model or hypothesis prior to the analysis. We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques and approaches like Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering and Association Rule Mining are unsupervised learning practices.

It is however extremely difficult and in a way pointless to talk about these techniques/algorithms/theories in terms of their accuracy or performance but however their logic can be studied and some of their building blocks can be talked of in terms of pros and cons.

## Classification

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithms require that classes be defined based on the data attribute value. It describes this class according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the et of parameters required for proper discrimination.

## Decision Trees

A decision tree is a flow chart like tree structure, where each node shows test on an attribute value, each branch represents the result of the test, and tree leaves represent classes. The drive model can be represented in different forms such as classification rules, decision tree, mathematical formula or neural networks. Decision trees are very similar to classification.

## Neural Networks

This technique was actualized by analyzing how the human nervous system functions. The area of neural networks probably belongs to the border line to and from A.I. A neural network is a collection of brain cells like processing units with weighted connection between the units. It contains many elements, called nodes which are joined. The joint between 2 nodes is weighted and by the measurement of this weight, training of the neural network is acheieved.

## Association rule

Association rule mining is the discovery of association relationships or correlation among a set of items. Association and correlation is used to find the frequent item set among large data sets. The main task of association rule mining is to find sets of binary variable that occur together frequently in a database.

## Clustering

Clustering as the name suggests is the process of grouping data into sets called clusters, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other cluster. Dissimilarities have been observed on the basis of attribute value describing the objects often distance used. How far data is from a particular cluster can be used to identify anomalies and data that are similar to a particular cluster.

# Popular data mining tools

Data is definitely priceless. But it is not a cake walk to analyze it as greater things come at a greater cost. With the exponential growth in data, there requires a process to extract meaningful information as conclude to useful insights. Now after going through all these techniques and approaches it must be obvious that implementation of these algorithms, techniques and approaches are no easy feat, which is why premade tools simplify some of the strenuous tasks that are involved with datamining. Here are some of the most used data mining tools as of 2017:

## Rapid Miner

Rapid Miner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis. It is one of the apex leading open source system for data mining.

The program is written entirely in Java. The program provides an option to try around with a huge number of arbitrarily nestable operators which are detailed in XML files and are made with graphical user interference of rapid miner.

## Oracle Data Mining

Market leading companies use Oracle data mining to maximize the potential of their data to make accurate predictions. The system works with a powerful data algorithm to target best customers.

Also, it identifies both anomalies and cross-selling opportunities and enables users to apply a different predictive model based on their need. Further, it customizes customer profiles in the desired way.

## Orange

This is a tool that I have used personally, Orange is an open source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for exploratory data analysis and interactive data visualization.

Orange is a component-based visual programming software package for data visualization, machine learning, data mining and data analysis. Orange components are called widgets and they range from simple data visualization, subset selection and pre-processing, to evaluation of learning algorithms and predictive modeling.

Visual programming in orange is performed through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration.

## Weka

Another personally used tool. Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software developed at the University of Waikato, New Zealand.

The program is written in Java. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling coupled with graphical user interface. Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection.

# The building blocks of most traditional & modern approaches/techniques

The Naïve Bayesian theory serve the foundation of being the building block to almost all traditional data mining approaches and techniques, while CART is presumably the core of nearly all the newest implementations of modern approaches/techniques. Here I have **illustrated** the 2 based on their pros and cons:

|  |  |  |
| --- | --- | --- |
| **Theory** | **Pros** | **Cons** |
| Naïve Bayesian | * It is easy to understand. * It is not sensitive to unknown features. * It handles streaming of data pretty moderately. * It can handle real and unreal values. * Is very popular, and nearly any traditional or modern algorithm is compatible with it. | * It assumes independence of features even when there is insufficient data to back it up. Therefore provides less accuracy. |
| CART | * It does not require preselection of variables. * It easily handles and identifies anomalies quickly. | * It is pretty unstable during the implementation of a decision tree. * It allows only splitting by one parameter. |

# Scopes of data mining

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities, as such the following are the 2 primary scopes of data mining:

1. **Automated prediction of trends and behaviors** - Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly.

A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings.

Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

1. **Automated discovery of previously unknown patterns** - Data mining tools sweep through databases and identify previously hidden patterns in one step.

An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed.

Alongside the primary scopes, some of the secondary scopes that I believe that data mining consists of are:

1. Data mining processes work in such a manner that it allows business to grow in an exponential rate.
2. It optimizes crawling large databases for valuable information within a short time and provides features classified under business intelligence.
3. It represents data in some logical order or patterns must be easily observable.
4. If hierarchies can be identified, then they must be visualized in a tree-shaped structure.
5. Introduces a genetic process of classification of different data sets to view the data at a glance.

# Data Mining algorithms & how they work

There exists a lot of data mining algorithms, and information that is available to the public eye is quite scarce, as such I will discuss only 4 well-known data mining algorithms which are C4.5, K-means, Apriori and EM.

**C4.5**: C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified. This is a supervised learning algorithm, since the training data set is labelled with classes.

C4.5 operates in 4 main steps, firstly it uses information gaining functions when generating the decision tree, then it uses a technique called single pass pruning to avoid over-fitting, then it establishes thresholds to any continuous data so they’ll be discrete and finally any incomplete data is handled separately.

**k-means**: k-means creates k groups from a set of objects so that the members of a group are more similar. It’s a popular cluster analysis technique for exploring a dataset.

Cluster analysis is the technique that k-means employs. Cluster analysis is a family of algorithms designed to form groups such that the group members are more similar versus non-group members. Clusters and groups are synonymous in the world of cluster analysis.

**Apriori:** The Apriori algorithm learns association rules and is applied to a database containing a large number of transactions. Association rule learning is a data mining technique for learning correlations and relations among variables in a database.

All apriori algorithms have a 3 step approach, Join, Prune and Repeat. Firstly during the join step, the whole database is scanned for 1-item sets, then those that satisfy support and confidence conditions move to the next step for 2-item sets, then the previous stages are repeated till the pre-defined level of item sets have been reached.

**EM:** In data mining specifically, Expectation Maximiazation is generally used as a clustering analysis algorithm like k-means for KDD. The EM algorithm iterates and optimizes the likelihood of seeing observed data while estimating the parameters of a statistical model with unobserved variables.

# k-means algorithm implementation

As a real-world scenario can’t properly captured into a course material like this, I will make logical and reasonable assumptions that will set up a case scenario and then achieve it.

This implementation will follow a 3 step procedure:

1. Initialization – “Centroids” will be randomly generated. A centroid is a value that will serve as the role of data in this implementation.
2. Relationship identification – K clusters are created by observing the nearest centroid for each given centroid.
3. Finalization – The centroid of the clusters will become the average of the clusters.

Let’s assume there is a bakery, this bakery delivers buns, breads and many other baked delicacies to several food companies in the country. We will be using the k-means algorithm to identify what type of food sells the most in a particular region of the country, as such the independent variable will be regions of the country and the dependent variable will be people that consume the food from this bakery. The regions and the people will obviously be given a numerical value to represent them.

I will use Python, and a several of its libraries like pandas, numpy and a library known as matplotlib to plot graphs to visualize the clusters in each stage of the implementation. The 3 types of food we will be checking for will be buns, breads and cakes represented by red, green and blue respectively.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

df = pd.DataFrame({

'x': [12, 20, 28, 18, 29, 33, 24, 45, 45, 52, 51, 52, 55, 53, 55, 61, 64, 69, 72],

'y': [39, 36, 30, 52, 54, 46, 55, 59, 63, 70, 66, 63, 58, 23, 14, 8, 19, 7, 24]

})

np.random.seed(200)

k = 3

centroids = {

i+1: [np.random.randint(0, 80), np.random.randint(0, 80)]

for i in range(k)

}

fig = plt.figure(*figsize*=(5, 5))

plt.scatter(df['x'], df['y'], *color*='k')

colmap = {1: 'r', 2: 'g', 3: 'b'}

for i in centroids.keys():

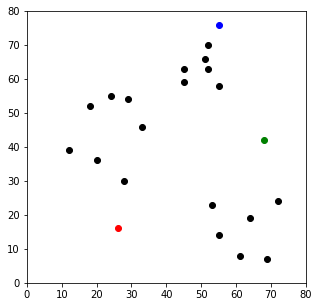
plt.scatter(\*centroids[i], *color*=colmap[i])

plt.xlim(0, 80)

plt.ylim(0, 80)

plt.show()

Firstly, the values for the X axis and Y axis are generated, then are scattered across the graph while also placing our cluster heads for buns, breads and cakes in RBG, which outputs the following graph.



As you can see, none of the values are in any order. The goal of initialization is to get values and scatter them, and it has been done perfectly.

Next, clusterification will be done by observing nearest centroids

*def* identification(*df*, *centroids*):

for i in centroids.keys():

df['distance\_from\_{}'.format(i)] = (

np.sqrt(

(df['x'] - centroids[i][0]) \*\* 2

+ (df['y'] - centroids[i][1]) \*\* 2

)

)

centroid\_distance\_cols = [

'distance\_from\_{}'.format(i) for i in centroids.keys()]

df['closest'] = df.loc[:, centroid\_distance\_cols].idxmin(*axis*=1)

df['closest'] = df['closest'].map(

*lambda* *x*: *int*(x.lstrip('distance\_from\_')))

df['color'] = df['closest'].map(*lambda* *x*: colmap[x])

return df

df = identification(df, centroids)

fig = plt.figure(*figsize*=(5, 5))

plt.scatter(df['x'], df['y'], *color*=df['color'], *alpha*=0.5, *edgecolor*='k')

for i in centroids.keys():

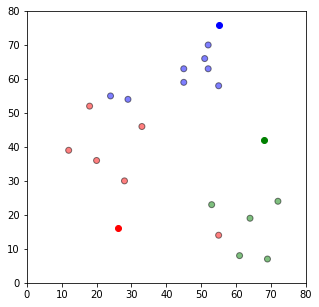
plt.scatter(\*centroids[i], *color*=colmap[i])

plt.xlim(0, 80)

plt.ylim(0, 80)

plt.show()

Which gives us the following output:



Now as you see, like values are color-identified and are much closer to their respectively related cluster. The next stage will focus on finalizing, which means that anomalies will be removed and if possible properly clustered and the cluster’s main node will become the average of that cluster.

*def* final(*k*):

for i in centroids.keys():

centroids[i][0] = np.mean(df[df['closest'] == i]['x'])

centroids[i][1] = np.mean(df[df['closest'] == i]['y'])

return k

centroids = final(centroids)

fig = plt.figure(*figsize*=(5, 5))

ax = plt.axes()

plt.scatter(df['x'], df['y'], *color*=df['color'], *alpha*=0.5, *edgecolor*='k')

for i in centroids.keys():

plt.scatter(\*centroids[i], *color*=colmap[i])

plt.xlim(0, 80)

plt.ylim(0, 80)

for i in old\_centroids.keys():

old\_x = old\_centroids[i][0]

old\_y = old\_centroids[i][1]

dx = (centroids[i][0] - old\_centroids[i][0]) \* 0.75

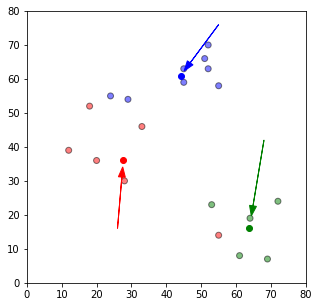
dy = (centroids[i][1] - old\_centroids[i][1]) \* 0.75

ax.arrow(old\_x, old\_y, dx, dy, *head\_width*=2,

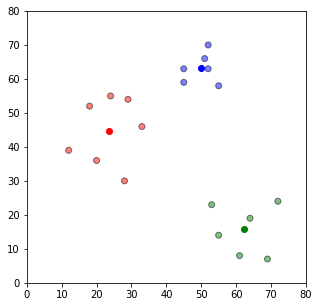
*head\_length*=3, *fc*=colmap[i], *ec*=colmap[i])

plt.show()

I’ve generated 3 arrows to show the movement of the values towards their respective cluster, this point on I will no longer show any code because there is a lot left, please refer to file called “k-means.py” in this folder to continue, comments have been added to help navigate the codebase. I advise using JetBrains PyCharm, Jupyter Notebook or Visual Studio Code with Python extensions to view the code for the best experience. This will be the output:



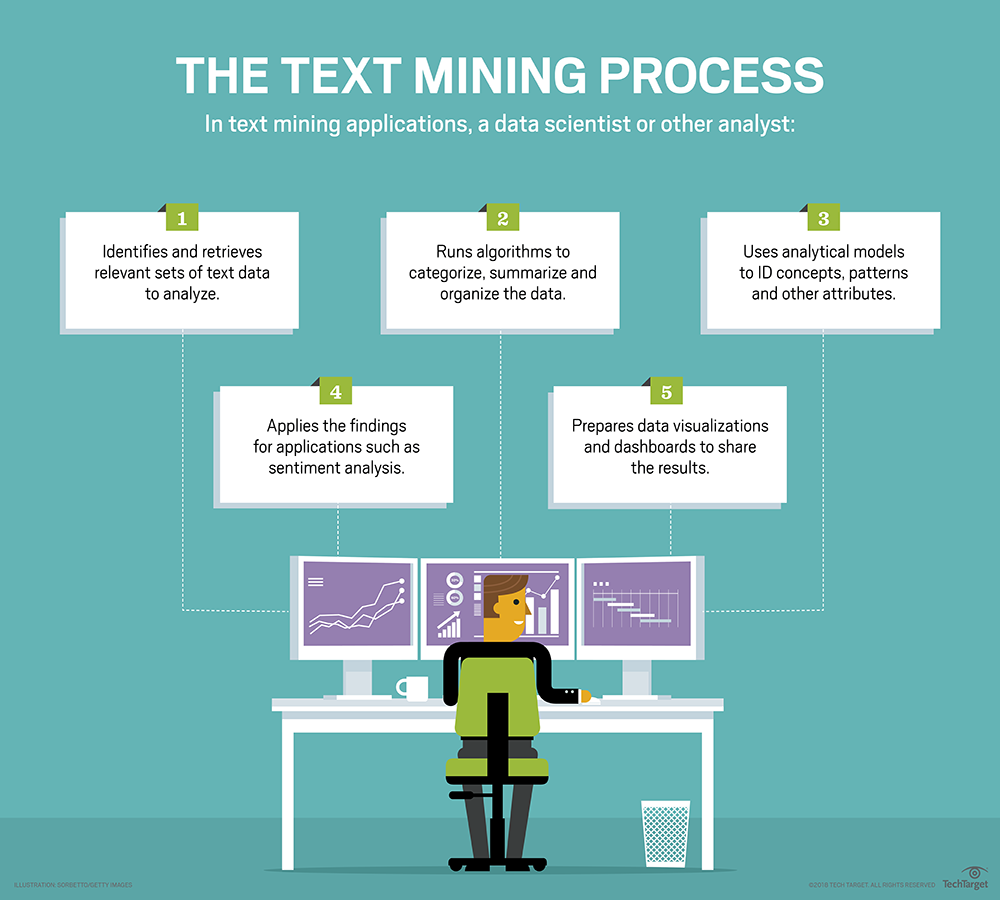
And then after complete finalization and running the entire codebase, the following will be the final output, all of the values are clustered properly, each with their main centroid that is the average of all of the values in that cluster. As you can see, the red cluster has 7 values, blue has 6 and green has 6 therefore with confidence you can say that buns are the most consumed food from the bakery.



# Text mining & how it works

Figure 1.1, whatis.com, Mrs. Illeva

**T**ext mining is the elaborative process of analyzing and searching through large amounts of text data that have no apparent structure using software and algorithms that can identify concepts, patterns, keywords and other attributes of the data. It is more often known as text analytics, but there is a debate as to the validity of this statement saying that text analytics is more of an application based scenario enabled using the text mining techniques to sort data sets.



However though, as the entirety of text analytics and text mining come to one result in the end and obeys one purpose, it is not inaccurate to say they’re related.

Text mining has become more famous among data scientists and other users due to the increase in the number of big data platforms and deep learning demand that can analyze massive sets of unstructured text.

Analyzing text helps organizations in a plenty of ways, to find valuable business insights in documents, emails, logs, surveys, social media posts, medical records and other forms of text.

Text mining capabilities are more and more introduced weak A.I. chat bots in the corporate industry and agents that companies usually outsource to provide automated responses to customers as part of the business’ marketing, sales or customer service operations.

# Text mining techniques, methods and approaches

Text mining is very similar to data mining but more with a scope based on text instead than structured data. One of the first steps in text mining process is to structure the data in some way so it can be put under qualitative and quantitative analysis.

This is done using the help of what’s known as natural language processing technology, which basically applies computational principles in order to parse and interpret data sets.

At a glance, text mining includes categorizing, clustering and tagging text; summarizing data sets and extracting information about things such as word occurrence rates and relationships between particular data entities. Analytical models are finally run to generate findings that can help business management, strategies or operational level actions.

## Text mining algorithms and how they work

The most commonly used text mining algorithms are those used for classification. Text mining is basically a task that when considering a pair of entities that occur in a particular sentence, the objective is to categorize the relations based on some predefined list.

Barely any other algorithms are used in text mining, when classification is used the company has to decide depending on a set of rules that should be applied so as to obtain the result they wish to obtain.

When rule-based classification is used, there are 2 main steps in the process of text mining; Named entity recognition and relation extraction.

In this process, named entity recognition is simply identifying any words that pertain to something that exists in the real world such as “YouTube” or “banana”, therefore its purpose is to extract these particular identities from the unstructured texts and give them a particular type for example “banana” will be given the type fruit.

The next step is relation extraction that basically revolves around the output of NER. Relation extraction is the process by which semantic relations are made between previously mentioned entities in text. So “banana” will be have a relation with “apple” because they’re both fruits.

Then a final resulting document of findings will be generated after the completion of named entity recognition and relation extraction.

Text mining as previously mentioned is used heavily in chat bots or bots basically nowadays, these bots act with no real frontal appearance, the final intended purpose of text mining in bots is usually to provide statistical information to users.

# Real world examples of text mining

For example, some big ecommerce sales websites are said to use text mining on their own social media portals to get a grasp of what customers like the most, text mining will be performed on the text in social media portals and then a findings report will be generated with the help of this finding report the management would usually offer discounts or put items that the customers wish to buy in sales.

In the case of a chat bot, these statistical data will be used to suggest customers with what to buy after a purchase or during search.

Another example of text mining comes in social media, although this example is not true for every social media portal it is certainly true for the giants like Facebook and twitter, these websites have a concept known as “tags”.

Tags are basically pointers that help named entity recognition, a feature-based classification is performed over their large bases of text and then a finding report Is generated, the relations between these tags are used to suggest users with hot topics, or other users who tend to talk about the same tags.

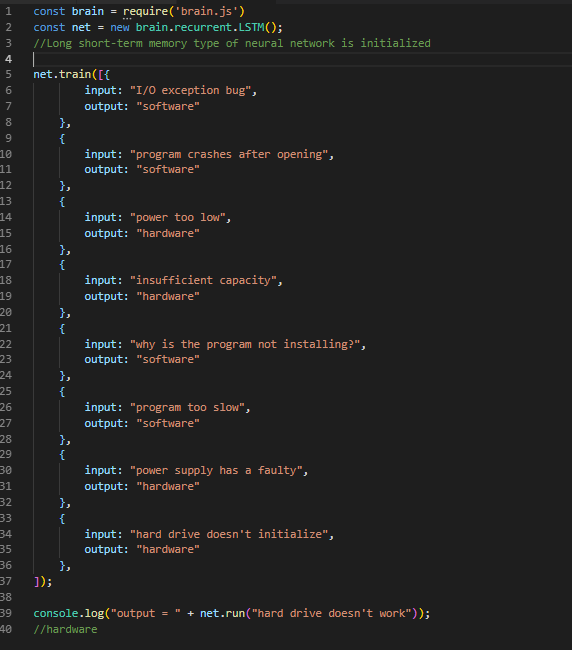
Please watch this video now; <https://youtu.be/xxqrIZyKKuk>

# Real world implementation

This implementation is to identify whether a given problem is hardware or a software fault. I will once again use JS and Node JS to write a text mining implementation using the help of a neural network.

Here I will train the neural network with various sentences be it questions, phrases or just words then the output will be introduced as hardware or software, when sufficient number of training data has been introduced, I will feed a problem to the neural network and attempt to draw a conclusion from it, whether the problem is hardware or software based.

## Code & Output



One may argue that the conclusion is accurate as the words in the problem exist in one of the inputs given to train the test data, although this is a valid question it has a simple answer: So do industrial text mining approaches.

The solid focus in text mining is to have a large data set that contains words that may be present in nearly any query made to the neural network, therefore helping to provide valid conclusions. Now as in this implementation the number of training data is low, this query may even give software as the conclusion, this is again due to the insufficient quantity of the data set.

# Graph mining

The task of graph mining is to extract patters (sub-graphs) of interest from graphs, that describe the underlying data and could be used further, e.g., for classification or clustering.

# Graph mining algorithms

Even though sub-graph isomorphism is a NP-complete problem, many graph mining tools for frequent sub-graph mining exist (like e.g., gSpan or GASTON) that can be applied to large databases (due to efficient candidate generation and unique canonical representations).

Graph mining has a vast number of applications, e.g. biological networks or web data. Chemical Informatics is another important application of graph mining: frequent sub-graph mining can yield structural alerts, i.e., structural sub-graphs that have a huge impact on the activity of chemical compounds (as used in Chemical informatics and Predictive Toxicology).

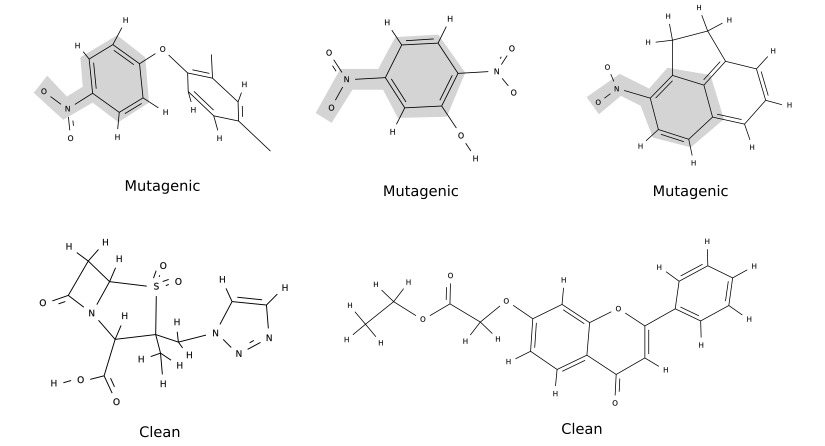
Graph mining unlike the previously mentioned types of mining is not extensively nor broadly used in the industry, mainly due to the lack of expertise. As such information regarding graph mining is very low.

# Real world examples of Graph Mining

Graph data mining is heavily used in organic chemistry within the industry, for example an organic compound like 2,4-dinitrophenylhydrazine (commonly called brady’s reagent) used to test for ketones and aldehydes or more rather the presence of the C = O bond, is produced artificially.

During this process the density of electrons have to be properly analyzed, so the structure of the reactants are graphed and tested against a large list of other reactant structures through the help of graph mining. By creating relations and establishing findings, we can successfully identify whether or not the reagent has been produced properly.

Also another example could be also in chemistry but based on medicinal fields during the analysis of organic compounds, in IR-spectroscopy. In this process a particular compound is allowed to absorb IR rays which results in the oscillation and changes in the bond lengths within the compound, by how much oscillation and received rays we can identify what type of bonds are present in the compound.



A graph is plotted during the analysis that originally makes no sense, through the help of graph mining however this senseless graph can be compared using classification and association techniques to identify the bonds and group them accordingly and producing an evaluation report.

This analysis can help identifying whether or not a compound is mutated or whether if it’s safe to be used in medicine.

Figure 1.2, dtai.cs.kuleuven.be, Dr. Jan Ramon, 30/11/2014

# Real world implementation

I have taken upon it myself to perform graph mining on a hypothetical data set that I visualize which will be containing concentration and other information about ingredients in whiskeys. I will use a neural network to perform graph mining and then obtain valid conclusions.

Graph mining in its essence is actually performing data mining techniques on a set of numerically expressible values, this example will attempt to mine the hypothetical CSV file and use the concentration of sulfur, sugar and alcohol in the whiskey to determine the quality of it.

We will assume that there are 1000 values in the CSV file, and first convert the data into JSON format and then feed it to the neural network. The conversion from CSV to JSON will be the data cleaning stage, as CSV lacks structure but JSON has structure.

The quality of the whine will be ranged from 0 to 1, where 0 is the worst and 1 being the best possible.

The implementation will be done using JS and a server runtime known as NodeJS. We will also use loadash, an external library to perform appropriate cleaning on data.

Below you will see the code used to create the neural network, then train it and finally obtain conclusions from it.

This is a real world example based in alcoholic beverages industry, at one point the industry found it increasingly difficult to maintain and identify quality alcoholic beverages therefore, they created large data sets containing concentrations, proportions and type of ingredients in them and performed data mining techniques like data mining.

Although this implementation is vastly smaller in size, it generally captures the concept in the modern industry today.

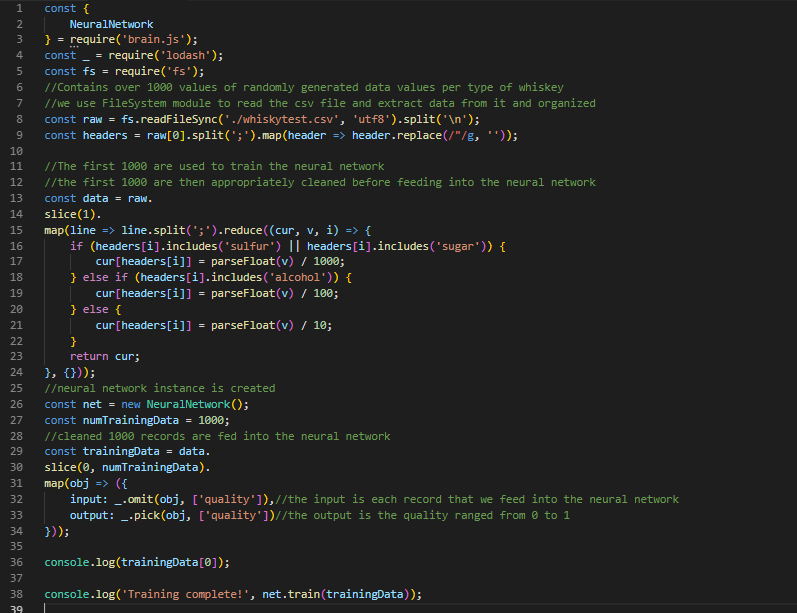
This is how the CSV file will look like:

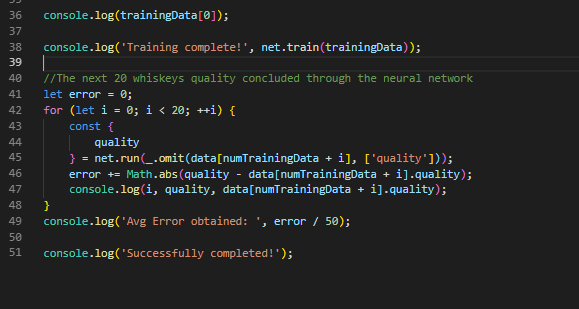
id;sulfur;sugar;alcohol

1;0.0436;0.01239;.6928

2;0.0299;0.00832;.0.4859

## Code





## Ouput

Here after using the neural network to draw conclusions you see 18th and 19th whiskeys being given 0.5 and 0.6 quality. But the average error that has been obtained is fairly high though, as such this cannot be accepted as a valid conclusion but this is entirely because of the lack of the number of training data I have used. Higher the training data, higher the quality.

5