Classification of Dataset using Data Mining and Machine Learning key words with Wordcloud

Name: Chandrashekar Akkenapally

Email:chandrashekar.akkenapally@mst.edu (mailto:chandrashekar.akkenapally@mst.edu)

StudentID:12589492

Course: CS 5402

Assignment: Programming Assignment 1

Date: 06/17/2022

Content Description

We are going to study the keywords majorly used in definitions of Data Mining and Machine Learning extracted from first 15 result sites when surfed their definitions in google and these are represented with the help of Wordcloud.

Data Collection

The extraction of data which defines "Data Mining" and "Machine Learning" is done from 15 websites of each search result and stored in csv file. After which we have done wordcloud using python to represent the maximum recurring keywords.

Data Description

Data Mining and Machine Learning definitions are stored in each csv file along with the different attributes:
* **Category:** Will specify the category i.e Data Mining or Machine Learning
  Example: Data Mining
* **Date:** The time when we extracted the data from website.
  Example:06/15/2022
* **Website:** To Identify the website name.
  Example: IBM
* **URL** Exact website link or URL from which the data is taken.
  Example:https://www.ibm.com/cloud/learn/data-mining
* **Selected Text Block:**Defines the meaning of the category taken from the specific websites.
  Example: Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets.

Data Import and Wrangling

* Data Mining and Machine Learning csv files are read using pandas.
* Importing all the required functions for data mining and analysis.
* Displaying the top 5 rows from the Data Mining and Machine Learning csv files.

In [1]:

```
import numpy as npy
import pandas as inputRead
import collections
import matplotlib.cm as cm
import matplotlib as matplt
from matplotlib import rcParams
import matplotlib.pyplot as plot
from wordcloud import WordCloud, STOPWORDS
```

In [3]:

```
dataMiningCollection = inputRead.read_csv(r"C:\Users\MYPC\Documents\Introduction to Data Mi
                   skip_blank_lines=True,na_filter=True,encoding='latin-1')
dataMiningCollection=dataMiningCollection.dropna()
dataMiningCollection.head() #head method displays the top 5 rows in Data Mining
```

Out[3]:

	Category	Date	Website	URL	Selecte Text Bloc
0	Data Mining	10:58 AM	www.sas.com	https://www.sas.com/en_us/insights/analytics/d...	Da mining tl process findi anomalie
1	Data Mining	10:59 AM	www.ibm.com	https://www.ibm.com/cloud/learn/data-mining	Da minin al known knowled discovery
2	Data Mining	11:00 AM	www.techtarget.com	https://www.techtarget.com/searchbusinessanaly...	Da mining tl process sorti through
3	Data Mining	11:01 AM	en.wikipedia.org	https://en.wikipedia.org/wiki/Data_mining	Da mining tl process extracti and d
4	Data Mining	11:03 AM	www.investopedia.com	https://www.investopedia.com/terms/d/dataminin...	Da mining is proce used compani to

In [6]:

```
machineLearningCollection = inputRead.read_csv(r"C:\Users\MYPC\Documents\Introduction to Da
                    skip_blank_lines=True,na_filter=True,encoding='latin-1')
machineLearningCollection=machineLearningCollection.dropna()
machineLearningCollection.head() #head method displays the top 5 rows in Machine Learning
```

Out[6]:

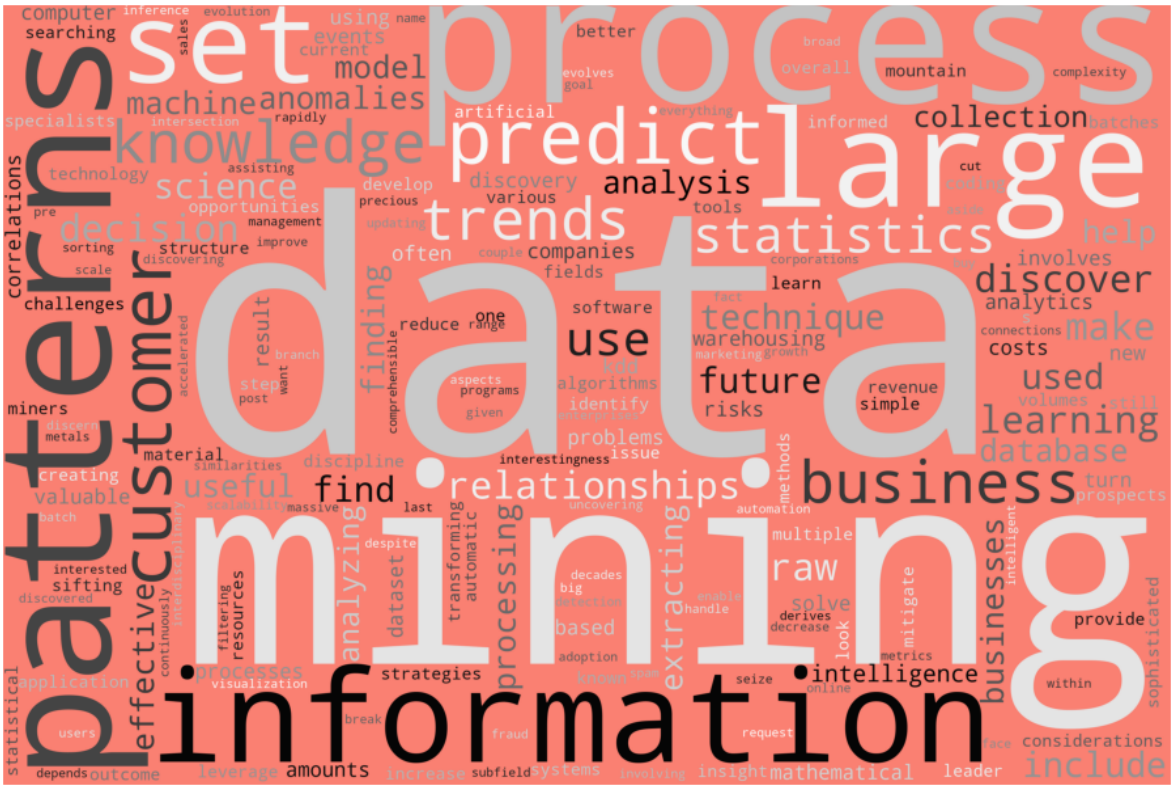
	Category	Date	Website	URL	Selected Text Block
0	Machine Learning	11:35 AM	www.ibm.com	https://www.ibm.com/cloud/learn/machine-learning	Machine learning is a branch of artificial int...
1	Machine Learning	11:37 AM	www.expert.ai	https://www.expert.ai/blog/machine-learning-de...	Machine learning is an application of AI that ...
2	Machine Learning	11:38 AM	en.wikipedia.org	https://en.wikipedia.org/wiki/Machine_learning	Machine learning (ML) is a field of inquiry de...
3	Machine Learning	11:39 AM	www.sas.com	https://www.sas.com/en_us/insights/analytics/m...	Machine learning is a method of data analysis ...
4	Machine Learning	11:41 AM	mitsloan.mit.edu	https://mitsloan.mit.edu/ideas-made-to-matter/...	Machine learning is a subfield of artificial i...

Exploratory Data Analysis:

* Wordcloud is build with the help of matplotlib function.
* Converting all the words to lowercased strings will help us to remove the deplucation.

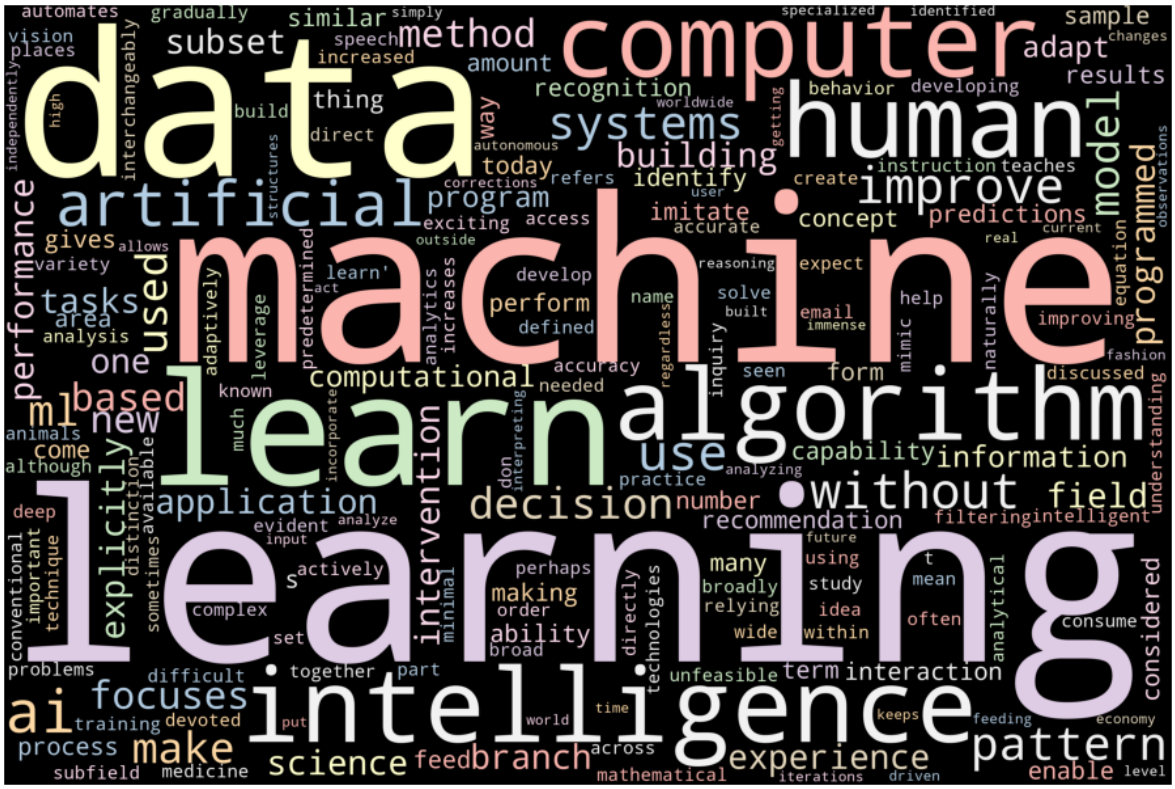
In [9]:

```
def plot_cloud(wordcloud):
    plot.figure(figsize=(16,12))
    plot.imshow(wordcloud)
    plot.axis("off");
stopwords = STOPWORDS
dataMiningDisplayText = ' '.join([text for text in dataMiningCollection['Selected Text Bloc
wordcloud = WordCloud(width= 3000, height = 2000, random_state=1,
                      background_color='salmon', colormap='gist_gray', collocations=False,
                      stopwords = STOPWORDS).generate(str(dataMiningDisplayText))
plot_cloud(wordcloud) #Data Mining wordcloud is displayed
```



In [10]:

```
def plot_cloud(wordcloud):
    plot.figure(figsize=(16,12))
    plot.imshow(wordcloud)
    plot.axis("off");

stopwords = STOPWORDS
machineLearningDisplayText = ' '.join([text for text in machineLearningCollection['Selected
wordcloud = WordCloud(width= 3000, height = 2000, random_state=1,
                      background_color='Black', colormap='Pastel1', collocations=False,
                      stopwords = STOPWORDS).generate(str(machineLearningDisplayText))
plot_cloud(wordcloud) # Machine Learning wordcloud is displayed
```



Mining or Analytics

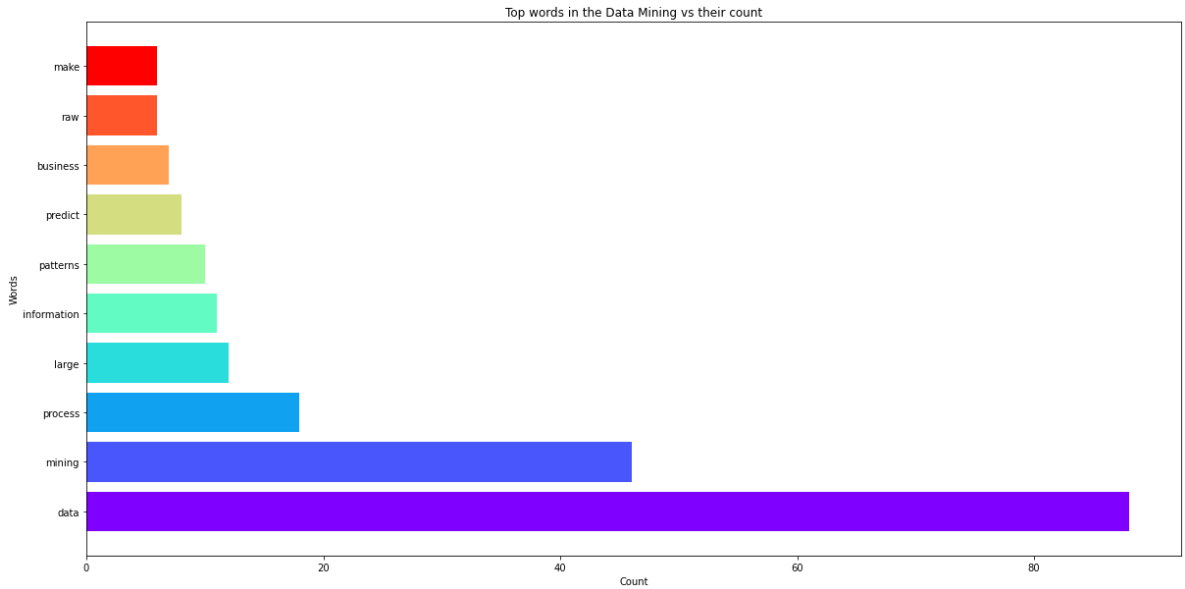
* Using the matplotlib function we will display the bar graph, where we can find the top 10 words used in the word cloud for both Data
  Mining and Artificial Intelligence.
* X-axis represents the count and Y-axis represents the words.

In [12]:

```
filtered_words = [word for word in dataMiningDisplayText.split() if word not in stopwords]
counted_words =collections.Counter(filtered_words)

words = []
counts = []
for letter, count in counted_words.most_common(10):
    words.append(letter)
    counts.append(count)
stopwords = STOPWORDS
colors = cm.rainbow(npy.linspace(0, 1, 10))
rcParams['figure.figsize'] = 20, 10

plot.title('Top words in the Data Mining vs their count')
plot.xlabel('Count')
plot.ylabel('Words')
plot.barh(words, counts, color=colors)
```

Out[12]:

```
<BarContainer object of 10 artists>
```

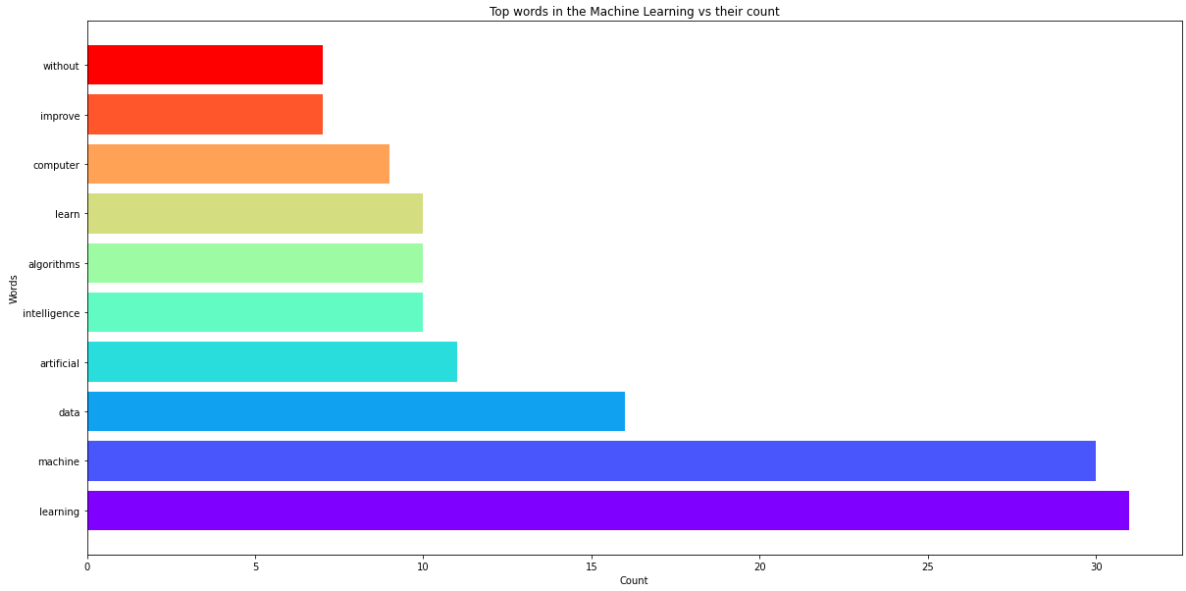


In [13]:

```
filtered_words = [word for word in machineLearningDisplayText.split() if word not in stopwo
counted_words =collections.Counter(filtered_words)

words = []
counts = []
for letter, count in counted_words.most_common(10):
    words.append(letter)
    counts.append(count)

colors = cm.rainbow(npy.linspace(0, 1, 10))
rcParams['figure.figsize'] = 20, 10

plot.title('Top words in the Machine Learning vs their count')
plot.xlabel('Count')
plot.ylabel('Words')
plot.barh(words, counts, color=colors)
```

Out[13]:

```
<BarContainer object of 10 artists>
```



Evaluation

By using the wordcloud and bargraphs we can easily understand the most used words in Data Mining and Machine Learning definition, which are taken from 15 different websites.

Results

* Word cloud gives us the Data visualization of the Data Mining and Machine Learning csv files.
* From the bargraph we can identify the top 10 words mostly used in the wordcloud.
* The word Data is used the most number of times in the Data wordcloud with count of 90, then comes with Mining,process
and information respectively.
* In the Machine Learning wordcloud the most commonly used word is ML with count greater thatn 30. The words
Machine Learning shares second postion,3rd & 4th postion words are data and artificial respectively.

References

1. Word cloud in python https://towardsdatascience.com/identify-top-topics-using-word-cloud-9c54bc84d911
2. Generating WordClouds in Python, link:https://www.datacamp.com/community/tutorials/wordcloud-python
3. The Jupyter Notebook Formatting Guide https://medium.com/analytics-vidhya/the-jupyter-notebook-formatting-guide-873ab39f765e
4. Pandas to read csv file https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html
5. Pandas tutorial, link: https://www.w3schools.com/python/pandas/default.asp