

```
##TEAM "RaAD" #Aditya Kumar #Rahul Patil
```

In [5]:

```
print("Team RaAD -- CultureScout HACKATHON\n Aditya Kumar\n Rahul Patil")
```

```
Team RaAD -- CultureScout HACKATHON
Aditya Kumar
Rahul Patil
```

In [6]:

```
!pip list
```

Package	Version
absl-py	1.4.0
alabaster	0.7.13
albumentations	1.2.1
altair	4.2.2
anyio	3.6.2
appdirs	1.4.4
argon2-cffi	21.3.0
argon2-cffi-bindings	21.2.0
arviz	0.15.1
astropy	5.2.2
astunparse	1.6.3
attrs	23.1.0
audioread	3.0.0
autograd	1.5
Babel	2.12.1
backcall	0.2.0
beautifulsoup4	4.11.2
' '	' ' '

## INSTALLING LIBRARIES

In [7]:

```
!pip install docx2txt
!pip install PyPDF2
!pip install -q transformers
!pip install multi_rake
!pip install git+https://github.com/LIAAD/yake
!pip install keybert

Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,)
https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheels/public/simple/)
Collecting docx2txt
  Downloading docx2txt-0.8.tar.gz (2.8 kB)
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: docx2txt
  Building wheel for docx2txt (setup.py) ... done
  Created wheel for docx2txt: filename=docx2txt-0.8-py3-none-any.whl size=3977 sha256=7bb864c46d2b502da9e4bb252a8a22d62299d1794925a2438bc95f4f94e1b1a5
  Stored in directory: /root/.cache/pip/wheels/22/58/cf/093d0a6c3ecfdfc5f6ddd5524043b88e59a9a199cb02352966
Successfully built docx2txt
Installing collected packages: docx2txt
Successfully installed docx2txt-0.8
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,)
https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheels/public/simple/)
```

## IMPORTING LIBRARIES

In [8]:

```
import re
import sys
import docx2txt
from PyPDF2 import PdfReader
import spacy
from spacy import displacy
from collections import Counter
from tqdm.autonotebook import tqdm
from transformers import pipeline
from transformers import AutoModelForSequenceClassification
from transformers import TFAutoModelForSequenceClassification
from transformers import AutoTokenizer
import numpy as np
from scipy.special import softmax
import csv
import urllib.request
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from transformers import AutoModelForQuestionAnswering, AutoTokenizer, pipeline
from multi_rake import Rake
import yake
from keybert import KeyBERT
```

```
<ipython-input-8-d97ff10af40c>:8: TqdmExperimentalWarning: Using `tqdm.autonotebook.tqdm` in notebook mode. Use `tqdm.tqdm` instead to force console mode (e.g. in jupyter console)
```

```
    from tqdm.autonotebook import tqdm
```

## TASK 1 INFORMATION EXTRACTION

####INCLUDES EXTRACTION OF ENTITIES LIKE EMAIL , PHONE\_NO, PERSON NAME, ORGANIZATION

In [9]:

```
#mount googledrive
from google.colab import drive
drive.mount('/content/drive/')
```

```
Mounted at /content/drive/
```

In [10]:

```
obtained_txt = docx2txt.process(r'/content/drive/MyDrive/Colab Notebooks/testNlp1.docx')
obtained_txt
```

Out[10]:

'1. Aditya kumar = aditya56678@gmail.com\n2. 9144249988\n3. ali khan\n4. 019-214-2344\nnamanGhosleLahsun@ac.in\nData mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their raw data into useful knowledge. However, despite the fact that technology continuously evolves to handle data at a large-scale, leaders still face challenges with scalability and automation. Data mining has improved organizational decision-making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two main purposes; they can either describe the target dataset or they can predict outcomes through the use of machine learning algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches. When combined with data analytics and visualization tools, like Apache Spark, delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within artificial intelligence only continue to expedite adoption across industries.'

In [11]:

```
#email and phone number format
email = re.compile(r'[a-zA-Z0-9-.]+@[a-zA-Z-.]*\.(com|ac.in|edu|net)')
phone_no = re.compile(r'(\d{3}[-\.\s]?\d{3}[-\.\s]?\d{4}|(\d{3})\s*\d{3}[-\.\s]?\d{4})')
```

Regex, which should also match 7 and 10-digit phone numbers that lack any hyphens, spaces or dots.

```
(\d{3}[-.\s]?\d{3}[-.\s]?\d{4})|((\d{3})\s*\d{3}[-.\s]?\d{4})|\d{3}[-.\s]?\d{4})
```

It matched the following values in RegexPal:

000-000-0000 USA

000 000 0000

000.000.0000

(000)000-0000

(000)000 0000

(000)000.0000

(000) 000-0000

(000) 000 0000

(000) 000.0000

000-0000

000 0000

000.0000

0000000

0000000000 INDIA

(000)0000000

In [12]:

```
#PRINTING ALL EMAILS AND PHONE NUMBERS FOUND IN DOCX FILE
```

```
emails = email.finditer(obtained_txt)
phone_nos = phone_no.finditer(obtained_txt)

for gotPhoneNums in phone_nos:
    print(gotPhoneNums.group(0))

for gotEmails in emails:
    print(gotEmails.group(0))
```

```
9144249988
019-214-2344
aditya56678@gmail.com
```

In [23]:

```
#LETS JUST SAY IF WE ARE WORKING WITH PDF FILE INSTEAD OF DOCX FILE
#TO PRINT EMAIL AND PHONE NUMBER

reader = PdfReader(r'/content/drive/MyDrive/Colab Notebooks/testNlp1.pdf')
num_pages = len(reader.pages)

counter = 0
while counter < num_pages:
    page = reader.pages[counter]
    text = page.extract_text()
    print(text)
    print("\n\n\n\n")
    emails_pdf = email.finditer(text)
    phone_nos_pdf = phone_no.finditer(text)

    print(f"Obtained PhoneNumbers at page {counter+1}\n")
    for gotPhoneNums in phone_nos_pdf:
        print(gotPhoneNums.group(0))

    print(f"\n\nObtained Emails at page {counter+1}\n")
    for gotEmails in emails_pdf:
        print(gotEmails.group(0))
    counter = counter+1
    print("\n\nGoing to next Page")

    print("\n\n")
```

1. Aditya kumar = aditya56678@gmail.com

2. 9144249988

3. ali khan

4. 019 -214-2344

amanGhosleLahsun@ac.in

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their raw data into useful knowledge. However, despite the fact that that technology continuously evolves to handle data at a large -scale, leaders still face challenges with scalability and automation. Data mining has improved organizational decision -making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two main purposes; they can either describe the target dataset or they can predict outcomes through the use of machine learning algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches. When combined with data analytics and visualization tools, like Apache Spark, delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within artificial

Obtained PhoneNumbers at page 1

9144249988

214-2344

Obtained Emails at page 1

aditya56678@gmail.com

Going to next Page

intelligence only continue to expedite adoption across industries.

Obtained PhoneNumbers at page 2

Obtained Emails at page 2

Going to next Page

In [24]:

```
#FOR IDENTIFICATION OF PERSONS , ORGANIZATION ETC ETC...
```

```
!pip install --upgrade spacy  
!pip install --upgrade spacy[cuda111,transformers]  
  
!python -m spacy download en_core_web_lg  
!python -m spacy download en_core_web_sm
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,),  
https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-pytho  
n.pkg.dev/colab-wheels/public/simple/)
```

```
Requirement already satisfied: spacy in /usr/local/lib/python3.10/dist-  
packages (3.5.2)
```

```
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/loca  
l/lib/python3.10/dist-packages (from spacy) (1.0.4)
```

```
Requirement already satisfied: pydantic!=1.8,!>1.8.1,<1.11.0,>=1.7.4 in  
/usr/local/lib/python3.10/dist-packages (from spacy) (1.10.7)
```

```
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/p  
ython3.10/dist-packages (from spacy) (1.1.1)
```

```
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-  
packages (from spacy) (3.1.2)
```

```
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/py  
thon3.10/dist-packages (from spacy) (2.0.7)
```

```
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python  
3.10/dist-packages (from spacy) (23.1)
```

```
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/  
lib/python3.10/dist-packages (from spacy) (1.0.9)
```

In [25]:

```
nlp = spacy.load("en_core_web_lg")  
nlp_sm = spacy.load("en_core_web_sm")  
  
nlp(obtained_txt)  
doc = nlp(obtained_txt)
```

In [26]:

```
#DISPLAY EACH TOKEN ACCORDING TO THEIR CLASSIFIED LABELS LIKE PERSON , ORGANIZATION , CARDINAL  
displacy.render(doc, style="ent", jupyter=True, manual=False)
```

1 CARDINAL . Aditya kumar PERSON = aditya56678@gmail.com

2 CARDINAL . 9144249988 CARDINAL

3 CARDINAL . ali khan PERSON

4 CARDINAL . 019 CARDINAL -214-2344

amanGhosleLahsun@ac.in

Data mining, also known as knowledge discovery in data ( KDD ORG ), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades DATE , assisting companies by transforming their raw data into useful knowledge. However, despite the fact that that technology continuously evolves to handle data at a large-scale, leaders still face challenges with scalability and automation. Data mining has improved organizational decision-making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two CARDINAL main purposes; they can either describe the target dataset or they can predict outcomes through the use of machine learning algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches. When combined with data analytics and visualization tools, like Apache Spark ORG , delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within artificial intelligence only continue to expedite adoption across industries.

In [29]:

```
#if anyone wants to what Label means what??
print(spacy.explain("CARDINAL"), "\n", spacy.explain("GPE"), "\n", spacy.explain("PERSON"), "
```

Numerals that do not fall under another type  
Countries, cities, states  
People, including fictional  
Absolute or relative dates or periods

In [30]:

```
just_text = obtained_txt
docs = list(tqdm(nlp.pipe(just_text), total=len(just_text)))

[(i.text, i.ent_iob_ + "-" + i.ent_type_) for i in doc[0:30]]
```

0% | 0/1440 [00:00<?, ?it/s]

Out[30]:

```
[('1', 'B-CARDINAL'),
 ('.', 'O-'),
 ('Aditya', 'B-PERSON'),
 ('kumar', 'I-PERSON'),
 ('=', 'O-'),
 ('aditya56678@gmail.com', 'O-'),
 ('\n\n', 'O-'),
 ('2', 'B-CARDINAL'),
 ('.', 'O-'),
 ('9144249988', 'B-CARDINAL'),
 ('\n\n', 'O-'),
 ('3', 'B-CARDINAL'),
 ('.', 'O-'),
 ('ali', 'B-PERSON'),
 ('khan', 'I-PERSON'),
 ('\n\n', 'O-'),
 ('4', 'B-CARDINAL'),
 ('.', 'O-'),
 ('019', 'B-CARDINAL'),
 ('-', 'O-'),
 ('214', 'O-'),
 ('-', 'O-'),
 ('2344', 'O-'),
 ('\n\n', 'O-'),
 ('amanGhosleLahsun@ac.in', 'O-'),
 ('\n\n', 'O-'),
 ('Data', 'O-'),
 ('mining', 'O-'),
 (',', 'O-'),
 ('also', 'O-')]
```

In [31]:

```
#PERSON IDENTIFICATION

all_persons = []
for d in doc:
    persons = [ent.text for ent in doc if ent.ent_type_ == "PERSON"]
    all_persons.extend(persons)

Counter(all_persons).most_common(15)
if len(all_persons) ==0:
    print("No Persons found in INPUT TEXT")
```

In [33]:

```
print(Counter(all_persons).most_common(15))

[('Aditya', 247), ('kumar', 247), ('ali', 247), ('khan', 247)]
```

In [34]:

```
#ORGANIZATION / COMPANY IDENTIFICATION

all_orgs = []
for d in doc:
    orgs = [ent.text for ent in doc if ent.ent_type_ == "ORG"]
    all_orgs.extend(orgs)

Counter(all_orgs).most_common(15)

if len(all_orgs) ==0:
    print("No organizations found in INPUT TEXT")
```

In [35]:

```
print(Counter(all_orgs).most_common(15))

[('KDD', 247), ('Apache', 247), ('Spark', 247)]
```

## SENTIMENT CLASSIFICATION

In [36]:

```
specific_model = pipeline(model="finiteautomata/bertweet-base-sentiment-analysis")
```

```
Downloading (...)lve/main/config.json: 0%| 0.00/949 [00:00<?, ?B/s]
```

```
Downloading pytorch_model.bin: 0%| 0.00/540M [00:00<?, ?B/s]
```

```
Downloading (...)okenizer_config.json: 0%| 0.00/338 [00:00<?, ?B/s]
```

```
Downloading (...)solve/main/vocab.txt: 0%| 0.00/843k [00:00<?, ?B/s]
```

```
Downloading (...)solve/main/bpe.codes: 0%| 0.00/1.08M [00:00<?, ?B/s]
```

```
Downloading (...)in/added_tokens.json: 0%| 0.00/22.0 [00:00<?, ?B/s]
```

```
Downloading (...)cial_tokens_map.json: 0%| 0.00/167 [00:00<?, ?B/s]
```

emoji is not installed, thus not converting emoticons or emojis into text.  
Install emoji: pip3 install emoji==0.6.0

In [39]:

```
user_txt_input = input("Enter any text\t")
```

Enter any text i love cats

In [40]:

```
#PRINTING MOST LIKELIHOOD ANSWER
```

```
data = [user_txt_input]
print(specific_model(data))
```

```
[{'label': 'POS', 'score': 0.9911149740219116}]
```

**'POS' --> POSITIVE**

**'NEG' --> NEGATIVE**

**'NEU' --> NEUTRAL**

In [42]:

```
# NOW IF WE WANT TO CLASSIFY SCORES FOR ALL EMOTIONAL_LABELS -->POSITIVE ,NEGATIVE, NEUTRAL

data_for_multi_classification = [user_txt_input]
data_txt = data_for_multi_classification[0]

final_result_for_plot = []
# Preprocess text (username and Link placeholders)
def preprocess(text):
    new_text = []

    for t in data_txt.split(" "):
        t = '@user' if t.startswith('@') and len(t) > 1 else t
        t = 'http' if t.startswith('http') else t
        new_text.append(t)
    return " ".join(new_text)

# Tasks:
# emoji, emotion, hate, irony, offensive, sentiment
# stance/abortion, stance/atheism, stance/climate, stance/feminist, stance/hillary

task='sentiment'
MODEL = f"cardiffnlp/twitter-roberta-base-{task}"

tokenizer = AutoTokenizer.from_pretrained(MODEL)

# download Label mapping
labels=[]
mapping_link = f"https://raw.githubusercontent.com/cardiffnlp/tweeteval/main/datasets/{task}.csv"
with urllib.request.urlopen(mapping_link) as f:
    html = f.read().decode('utf-8').split("\n")
    csvreader = csv.reader(html, delimiter='\t')
labels = [row[1] for row in csvreader if len(row) > 1]

# PT
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
#model.save_pretrained(MODEL)

#text = "Good night 😊"
text = preprocess(data_txt)
encoded_input = tokenizer(text, return_tensors='pt')
output = model(**encoded_input)
scores = output[0][0].detach().numpy()
scores = softmax(scores)

ranking = np.argsort(scores)
ranking = ranking[::-1]
for i in range(scores.shape[0]):
    l = labels[ranking[i]]
    s = scores[ranking[i]]
    print(f"{i+1}) {l} {np.round(float(s), 4)}")

1) positive 0.9773
2) neutral 0.0189
3) negative 0.0038
```

# VISUALIZATION

In [43]:

```
for i in range(3):
    print(labels[ranking[i]], scores[ranking[i]])
```

```
positive 0.9773393
neutral 0.018877784
negative 0.003782946
```

In [44]:

```
plt.style.use('seaborn')
```

```
<ipython-input-44-4a43041a1d49>:1: MatplotlibDeprecationWarning: The seabo
rn styles shipped by Matplotlib are deprecated since 3.6, as they no longe
r correspond to the styles shipped by seaborn. However, they will remain a
vailable as 'seaborn-v0_8-<style>'. Alternatively, directly use the seabor
n API instead.
```

```
plt.style.use('seaborn')
```

In [45]:

```
#bar plot
x_label = []
y_scores = []

for i in range(3):
    x_label.append(labels[ranking[i]])
    y_scores.append(scores[ranking[i]])

print(x_label)
print(y_scores)

plt.bar(x_label,y_scores,color='orange')

#adjust xticks
plt.xticks(rotation=70)

#label x-axis
plt.xlabel('Emotion Classes')

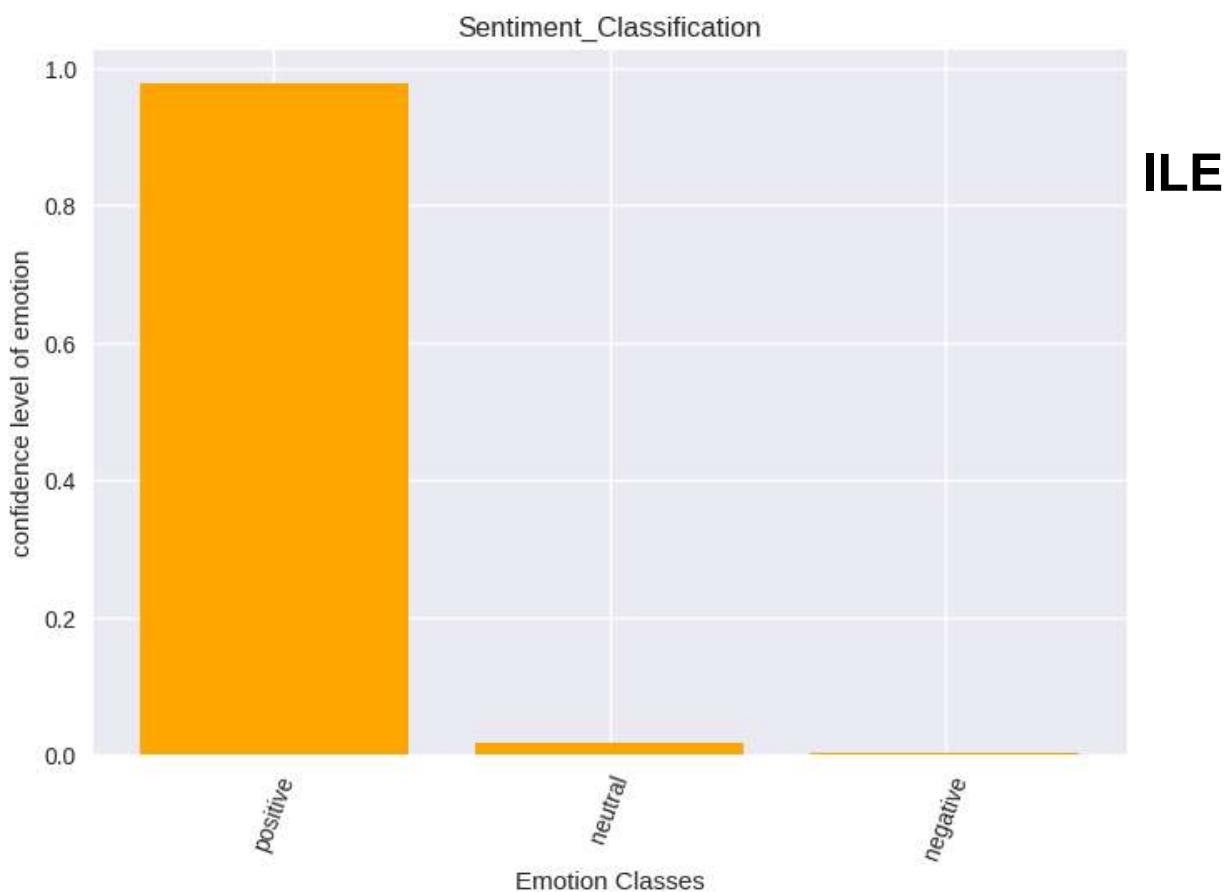
#label y-axis
plt.ylabel('confidence level of emotion')

#label the plot
plt.title('Sentiment_Classification')

#save plot
#plt.savefig('matplotlib_plotting_7.png',dpi=300,bbox_inches='tight')

#display plot
plt.show()

['positive', 'neutral', 'negative']
[0.9773393, 0.018877784, 0.003782946]
```



In [49]:

```
!python -m spacy download en_core_web_trf
```

```
2023-05-07 14:37:28.257786: W tensorflow/compiler/tf2tensorrt/utils/py_uti
ls.cc:38] TF-TRT Warning: Could not find TensorRT
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple/) ht
tps://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pk
g.dev/colab-wheels/public/simple/)
Collecting en-core-web-trf==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en\_core\_web\_trf-3.5.0/en\_core\_web\_trf-3.5.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en\_core\_web\_trf-3.5.0/en\_c
ore\_web\_trf-3.5.0-py3-none-any.whl) (460.3 MB)
    460.3/460.3 MB 2.8 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /usr/local/lib/python3.10/dist-packages (from en-core-web-trf==3.5.0) (3.5.2)
Requirement already satisfied: spacy-transformers<1.3.0,>=1.2.0.dev0 in /u
sr/local/lib/python3.10/dist-packages (from en-core-web-trf==3.5.0) (1.2.
3)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/p
ython3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0)
(3.3.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (2.
0.7)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /u
sr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-w
eb-trf==3.5.0) (1.10.7)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (4.
65.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dis
t-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (67.7.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-pa
ckages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (3.1.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/l
ib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.
5.0) (3.0.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/pyth
on3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0)
(1.1.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/p
ython3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0)
(2.0.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.1
0/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (23.1)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/li
b/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.
5.0) (1.0.9)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/p
ython3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0)
(2.27.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/pyt
hon3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0)
(3.0.8)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/pytho
n3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (8.
1.9)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/pytho
n3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (2.
4.6)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /usr/local/lib/pytho
n3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (0.
```

```
7.0)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (1.22.4)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (6.3.0)
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (0.10.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (1.0.4)
Requirement already satisfied: spacy-alignments<1.0.0,>=0.7.2 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (0.9.0)
Requirement already satisfied: transformers<4.29.0,>=3.4.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (4.28.1)
Requirement already satisfied: torch>=1.8.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (2.0.0+cu118)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!>1.8.1,<1.11.0,>=1.7.4->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (4.5.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (1.26.15)
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (3.4)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (0.0.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (0.7.9)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (1.11.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (3.12.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (3.1)
Requirement already satisfied: triton==2.0.0 in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (2.0.0)
Requirement already satisfied: cmake in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (3.25.2)
Requirement already satisfied: lit in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (16.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers<
```

```
1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (2022.10.31)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (0.13.3)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (6.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (0.14.1)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (8.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-web-trf==3.5.0) (2.1.2)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.11.0->transformers<4.29.0,>=3.4.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (2023.4.0)
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.8.0->spacy-transformers<1.3.0,>=1.2.0.dev0->en-core-web-trf==3.5.0) (1.3.0)
```

✓ Download and installation successful

You can now load the package via `spacy.load('en_core_web_trf')`

In [53]:

```
!pip install spacy-transformers
import spacy_transformers
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,) https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheels/public/simple/)
Requirement already satisfied: spacy-transformers in /usr/local/lib/python3.10/dist-packages (1.2.3)
Requirement already satisfied: transformers<4.29.0,>=3.4.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (4.28.1)
Requirement already satisfied: spacy<4.0.0,>=3.5.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (3.5.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (2.4.6)
Requirement already satisfied: torch>=1.8.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (2.0.0+cu118)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (1.22.4)
Requirement already satisfied: spacy-alignments<1.0.0,>=0.7.2 in /usr/local/lib/python3.10/dist-packages (from spacy-transformers) (0.9.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (2.0.7)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (1.0.9)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (4.65.0)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (3.3.0)
Requirement already satisfied: pydantic!=1.8,!>=1.8.1,<1.11.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (1.10.7)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (0.7.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (3.0.8)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (8.1.9)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (67.7.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (1.0.4)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (1.1.1)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (3.1.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (2.27.1)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (23.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (2.0.8)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers)
```

## (6.3.0)

Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.5.0->spacy-transformers) (0.10.1)  
Requirement already satisfied: triton==2.0.0 in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers) (2.0.0)  
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers) (4.5.0)  
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers) (3.1)  
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers) (3.12.0)  
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.8.0->spacy-transformers) (1.11.1)  
Requirement already satisfied: cmake in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch>=1.8.0->spacy-transformers) (3.25.2)  
Requirement already satisfied: lit in /usr/local/lib/python3.10/dist-packages (from triton==2.0.0->torch>=1.8.0->spacy-transformers) (16.0.2)  
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers) (0.13.3)  
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers) (6.0)  
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers) (0.14.1)  
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers<4.29.0,>=3.4.0->spacy-transformers) (2022.10.31)  
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.11.0->transformers<4.29.0,>=3.4.0->spacy-transformers) (2023.4.0)  
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.5.0->spacy-transformers) (2.0.12)  
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.5.0->spacy-transformers) (3.4)  
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.5.0->spacy-transformers) (1.26.15)  
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.5.0->spacy-transformers) (2022.12.7)  
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<4.0.0,>=3.5.0->spacy-transformers) (0.7.9)  
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<4.0.0,>=3.5.0->spacy-transformers) (0.0.4)  
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.8.0,>=0.3.0->spacy<4.0.0,>=3.5.0->spacy-transformers) (8.1.3)  
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<4.0.0,>=3.5.0->spacy-transformers) (2.1.2)  
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.8.0->spacy-transformers) (1.3.0)

In [54]:

```
nlp_trf = spacy.load("en_core_web_trf")

doc = nlp_trf(obtained_txt)
displacy.render(doc, style="ent", jupyter=True)
```

1. Aditya kumar PERSON = aditya56678@gmail.com

2. 9144249988

3. ali khan PERSON

4. 019-214-2344

amanGhosleLahsun@ac.in

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades DATE , assisting companies by transforming their raw data into useful knowledge. However, despite the fact that that technology continuously evolves to handle data at a large-scale, leaders still face challenges with scalability and automation. Data mining has improved organizational decision-making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two CARDINAL main purposes; they can either describe the target dataset or they can predict outcomes through the use of machine learning algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches. When combined with data analytics and visualization tools, like Apache Spark PRODUCT , delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within artificial intelligence only continue to expedite adoption across industries.

In [55]:

#DISPLAY GRAMMATICAL RELATIONS BETWEEN DIFFERENT TOKEN INSIDE CORPUS

```
doc = nlp(obtained_txt)
sent = list(doc.sents)[1]
displacy.render(sent, style="dep", jupyter=True)
```

Aditya PROPN kumar PROPN = PUNCT aditya56678@gmail.com X SPACE 2. NUM 9144249988 NUM SPACE 3. NUM ali PROPN khan PROPN SPACE 4. NUM 019- NUM 214- NUM 2344 NUM SPACE amanGhosleLahsun@ac.in NOUN SPACE Data NOUN mining, NOUN also ADV known VERB as ADP knowledge NOUN discovery NOUN in ADP data ( NOUN KDD), PROPN is AUX the DET process NOUN of ADP uncovering VERB patterns NOUN and CCONJ other ADJ valuable ADJ information NOUN from ADP large ADJ data NOUN sets. NOUN compound nsubj punct appos dep appos nummod dep appos compound nsubj dep appos nummod prep prep dep csubj dep compound dobj advmod acl prep compound pobj prep pobj appos det attr prep pcomp dobj cc amod amod conj prep amod compound pobj

In [56]:

```
display(sent.text)
```

'Aditya kumar = aditya56678@gmail.com\n\n2. 9144249988\n\n3. ali khan\n\n4. 019-214-2344\nnamanGhosleLahsun@ac.in\nData mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets.'

In [57]:

```
from transformers import AutoModelForQuestionAnswering, AutoTokenizer, pipeline
```

```
model_name = "deepset/roberta-base-squad2"
```

```
hugg = pipeline('question-answering', model=model_name, tokenizer=model_name)
```

Downloading (...)lve/main/config.json: 0%| 0.00/571 [00:00<?, ?B/s]

Downloading pytorch\_model.bin: 0%| 0.00/496M [00:00<?, ?B/s]

Downloading (...)okenizer\_config.json: 0%| 0.00/79.0 [00:00<?, ?B/s]

Downloading (...)olve/main/vocab.json: 0%| 0.00/899k [00:00<?, ?B/s]

Downloading (...)olve/main/merges.txt: 0%| 0.00/456k [00:00<?, ?B/s]

Downloading (...)cial\_tokens\_map.json: 0%| 0.00/772 [00:00<?, ?B/s]

In [58]:

```
#if User want to question
user_question = input("Enter your query\t")
```

Enter your query what is data mining

In [62]:

```
#DISPLAYING ANSWERS WITH REFERENCE TO THE CORPUS PRESENT IN FILE
```

```
QA_input = {
    'question': user_question,
    'context': sent.text
}
res = hugg(QA_input)

print(res) #result with Labels -->Raw answering

print("\n\n")
final_ans= res['answer']
print("ANSWER IS: ",final_ans) #Printing focused answer
```

{'score': 0.26368680596351624, 'start': 172, 'end': 258, 'answer': 'the process of uncovering patterns and other valuable information from large data sets'}

ANSWER IS: the process of uncovering patterns and other valuable information from large data sets

## CONTEXT AND KEYWORD IDENTIFICATION

In [67]:

```
obtained_txt2 = "Data mining, also known as knowledge discovery in data (KDD), is the pr
```

In [68]:

```
from multi_rake import Rake
rake = Rake()
keywords = rake.apply(obtained_txt2)

#all keywords related to this context on one list
print(keywords[:10])
```

[('technology continuously evolves', 9.0), ('improved organizational decision-making', 9.0), ('machine learning algorithms', 9.0), ('extracting relevant insights', 9.0), ('large data sets', 8.285714285714285), ('data warehousing technology', 8.285714285714285), ('data mining techniques', 7.685714285714285), ('insightful data analyses', 7.285714285714286), ('data mining', 4.685714285714285), ('big data', 4.285714285714286)]

In [69]:

```
#most relevant context to this corpus would be
for i in range(10):
    print(keywords[i])
    print("\n")
```

('technology continuously evolves', 9.0)

('improved organizational decision-making', 9.0)

('machine learning algorithms', 9.0)

('extracting relevant insights', 9.0)

('large data sets', 8.285714285714285)

('data warehousing technology', 8.285714285714285)

('data mining techniques', 7.685714285714285)

('insightful data analyses', 7.285714285714286)

('data mining', 4.685714285714285)

('big data', 4.285714285714286)

In [70]:

```
# ANOTHER METHOD USING RAKE FOR CONTEXT AND KEYWORD IDENTIFICATION
```

```
kw_extractor = yake.KeywordExtractor(top=10, stopwords=None)
keywords = kw_extractor.extract_keywords(obtained_txt2)
for kw, v in keywords:
    print("Keyphrase: ", kw, ": score", v)
```

Keyphrase: large data sets : score 0.012354533729270522  
Keyphrase: data mining techniques : score 0.021983002296349266  
Keyphrase: Data mining : score 0.022626926758312377  
Keyphrase: Data : score 0.026616928780721354  
Keyphrase: process of uncovering : score 0.0359885614189422  
Keyphrase: uncovering patterns : score 0.0359885614189422  
Keyphrase: mining : score 0.04797174582746027  
Keyphrase: knowledge discovery : score 0.05105873592824253  
Keyphrase: mining techniques : score 0.05267880670273334  
Keyphrase: KDD : score 0.05732400781576302

In [71]:

#ANOTHER METHOD USING KEYBERT

kw\_model = KeyBERT(model='all-mpnet-base-v2')

```
Downloading (...)a8e1d/.gitattributes:  0%|          | 0.00/1.18k [00:00<?, ?B/s]
Downloading (...)_Pooling/config.json:  0%|          | 0.00/190 [00:00<?, ?B/s]
Downloading (...)b20bca8e1d/README.md:  0%|          | 0.00/10.6k [00:00<?, ?B/s]
Downloading (...)0bca8e1d/config.json:  0%|          | 0.00/571 [00:00<?, ?B/s]
Downloading (...)ce_transformers.json:  0%|          | 0.00/116 [00:00<?, ?B/s]
Downloading (...)e1d/data_config.json:  0%|          | 0.00/39.3k [00:00<?, ?B/s]
Downloading pytorch_model.bin:  0%|          | 0.00/438M [00:00<?, ?B/s]
Downloading (...)nce_bert_config.json:  0%|          | 0.00/53.0 [00:00<?, ?B/s]
Downloading (...)cial_tokens_map.json:  0%|          | 0.00/239 [00:00<?, ?B/s]
Downloading (...)a8e1d/tokenizer.json:  0%|          | 0.00/466k [00:00<?, ?B/s]
Downloading (...)okenizer_config.json:  0%|          | 0.00/363 [00:00<?, ?B/s]
Downloading (...)8e1d/train_script.py:  0%|          | 0.00/13.1k [00:00<?, ?B/s]
Downloading (...)b20bca8e1d/vocab.txt:  0%|          | 0.00/232k [00:00<?, ?B/s]
Downloading (...)bca8e1d/modules.json:  0%|          | 0.00/349 [00:00<?, ?B/s]
```

In [72]:

```
keywords = kw_model.extract_keywords(obtained_txt2,  
                                     keyphrase_ngram_range=(1, 2),  
                                     stop_words='english',  
                                     highlight=False,  
                                     top_n=10)  
  
keywords_list= list(dict(keywords).keys())  
  
print(keywords_list)
```

```
['data mining', 'data analytics', 'big data', 'knowledge discovery', 'data  
warehousing', 'insightful data', 'analytics', 'uncovering patterns', 'mach  
ine learning', 'data adoption']
```

In [73]:

```
key_list = list(dict(keywords).keys())  
value_list= list(dict(keywords).values())  
  
for i in range(len(key_list)):  
    print(f"Found keyword \t{key_list[i]}\t with confidence value of \t{value_list[i]}")
```

```
Found keyword data mining      with confidence value of      0.6293  
Found keyword data analytics   with confidence value of      0.5871  
Found keyword big data        with confidence value of      0.5553  
Found keyword knowledge discovery with confidence value of  0.  
5498  
Found keyword data warehousing with confidence value of      0.  
5021  
Found keyword insightful data  with confidence value of      0.4971  
Found keyword analytics       with confidence value of      0.4833  
Found keyword uncovering patterns with confidence value of  0.  
483  
Found keyword machine learning with confidence value of      0.  
4786  
Found keyword data adoption    with confidence value of      0.4667
```