

ASSIGNMENT 1

17/01/24

NAME : SHRESTH SONKAR
REGNO : 20214272
GROUP : CS6D
TOPIC : DATA MINING LAB
CODE : CS-16202

```
# Q1 : WAP to obtain a sorted list of words with their  
# counts & make sure that all words are lower-cased &  
# contain only letters from a-z.
```

```
import string
```

```
text = open("/content/drive/MyDrive/Colab/24-01-17/  
data.txt", "r")  
d = dict()
```

```
for line in text:  
    line = line.strip()  
    line = line.lower()  
    line = line.translate(line.maketrans("", "",  
string.punctuation))  
    words = line.split(" ")
```

```
        for word in words:  
            if word in d:  
                d[word] = d[word] + 1  
            else:  
                d[word] = 1
```

```
for key in sorted(d.keys()):  
    print(key, ":", d[key])
```

➔ ~/desktop/cse/ASSGN/sem6/mine/2024-01-17 \$ python3 q1.py

```
a : 2
actionable : 1
already : 1
although : 1
analysis : 1
and : 7
applications : 1
as : 1
banking : 1
business : 1
can : 1
companies : 1
comprehensible : 1
concerned : 1
crucial : 1
data : 9
databases : 1
decisions : 1
examples : 1
extracting : 1
find : 1
finding : 1
focus : 1
for : 2
from : 1
have : 1
hidden : 2
huge : 1
in : 4
industries : 1
information : 2
insurance : 1
investment : 1
is : 6
it : 2
known : 1
large : 1
lists : 1
made : 1
make : 1
medicine : 1
mining : 6
new : 1
number : 1
of : 8
patterns : 1
paybacks : 1
previously : 1
process : 1
provide : 1
relationships : 1
relatively : 1
retailmarketing : 1
sets : 1
significant : 1
software : 1
still : 1
table : 1
techniques : 1
technology : 1
that : 1
the : 4
to : 2
unexpected : 2
unknown : 1
use : 1
used : 1
using : 1
valid : 1
warehousing : 1
who : 1
with : 1
```

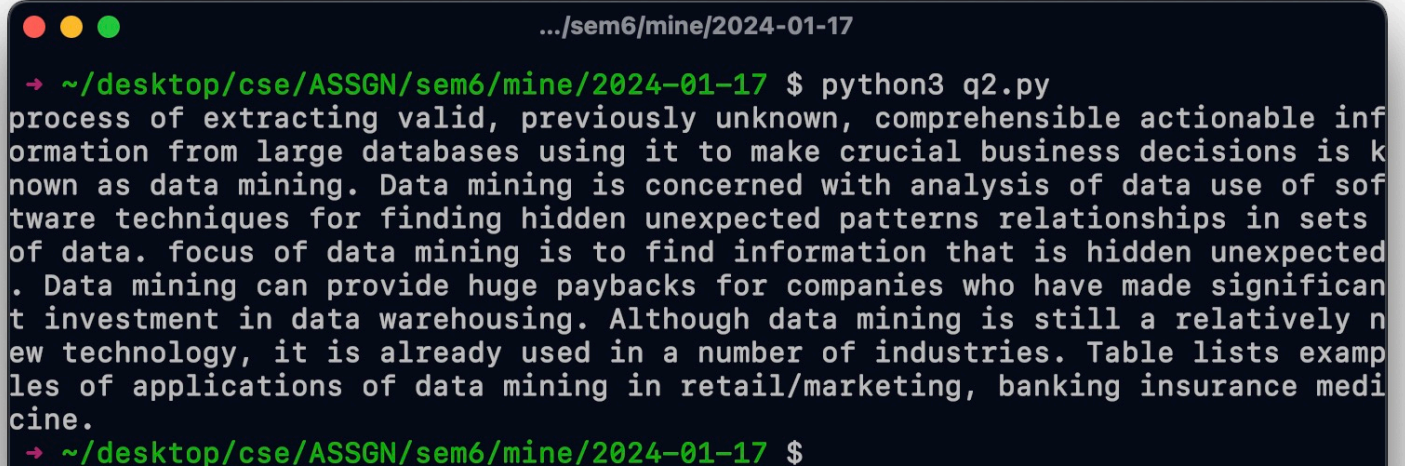
➔ ~/desktop/cse/ASSGN/sem6/mine/2024-01-17 \$

#Q2 : WAP to pre-process the data set by removing the most common words (stopwords). For English language some examples are "the", "and", "if", "which", and "on".

```
stpwrld_lst = ["the", "and", "if", "which", "on"]

with open("/content/drive/MyDrive/Colab/24-01-17/
data.txt", "r") as file:
    txt = file.read()

words = txt.split()
fltr_txt = [word for word in words if word.lower() not
in stpwrld_lst]
fltr_txt = ' '.join(fltr_txt)
print(fltr_txt)
```



```
.../sem6/mine/2024-01-17

→ ~/desktop/cse/ASSGN/sem6/mine/2024-01-17 $ python3 q2.py
process of extracting valid, previously unknown, comprehensible actionable inf
ormation from large databases using it to make crucial business decisions is k
nown as data mining. Data mining is concerned with analysis of data use of sof
tware techniques for finding hidden unexpected patterns relationships in sets
of data. focus of data mining is to find information that is hidden unexpected
. Data mining can provide huge paybacks for companies who have made significan
t investment in data warehousing. Although data mining is still a relatively n
ew technology, it is already used in a number of industries. Table lists examp
les of applications of data mining in retail/marketing, banking insurance medi
cine.
→ ~/desktop/cse/ASSGN/sem6/mine/2024-01-17 $
```