Analyzing mygov.in suggestions (mann-ki-baat)

Implementation Process

Data Mining using beautiful soup

Storing the data in a txt file

Nltk implementation (preprocessing)

Storing the pre processed data

Clustering Kmeans Algorithm

Classification based on likes

Data Mining using Beautiful Soup

Beautiful Soup is an application framework for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like data mining, information processing or historical archival.

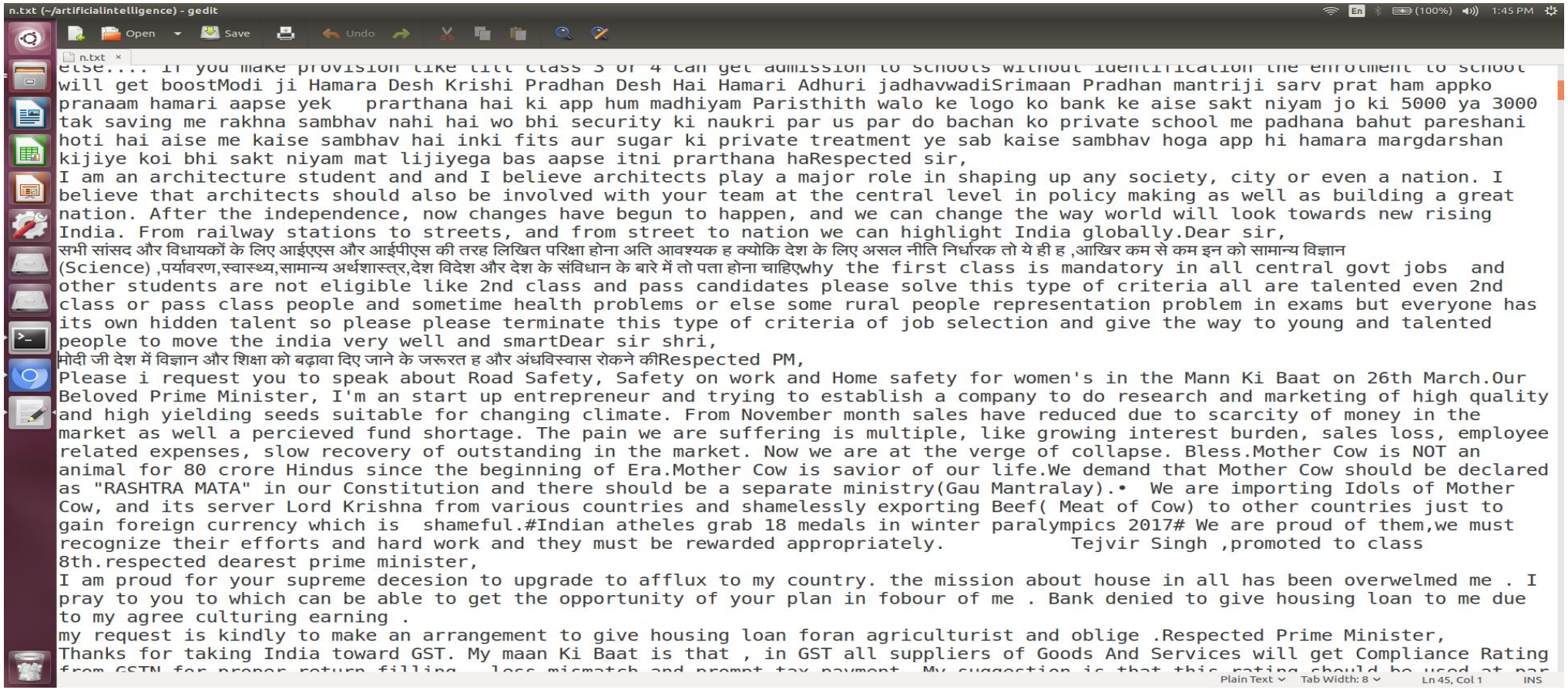
- It was originally designed for web scraping, it can also be used to extract data using APIs (such as Amazon Associates Web Services) or as a general purpose web crawler.
- Command to install scrappy in Linux :

```
$ pip install scrapy
```

Where the data came from?


```
<div class="comment_body"><p>
        Sir my humble submission is that please ask public not to man handle doctors because
they work in a very delicate situation, to save a patient is not always in his hand. The incidents
of manhandling doctors is increasing day by day and it's becoming very difficult to work in
these situatons. Majority are not Opting for medical profession, it will create a crisis in medical
field.In foreign no body can dare to manhandle a doctor, nurse, ambulance worker else he will
be behind bars for 14 years.</p>
        </div>
    </div>
    <div class="comment_extra_links">
     <div class="voting_wrap">
```

.txt file formed



n.txt (~/artificialintelligence) - gedit

Open Save Undo

n.txt

etse.... if you make provision like till class 3 or 4 can get admission to schools without identification the enrolment to school will get boostModi ji Hamara Desh Krishi Pradhan Desh Hai Hamari Adhuri jadhavwadiSrimaan Pradhan mantriji sarv prat ham appko pranaam hamari aapse yek prarthana hai ki app hum madhiyam Paristhith walo ke logo ko bank ke aise sakt niyam jo ki 5000 ya 3000 tak saving me rakhna sambhav nahi hai wo bhi security ki naukri par us par do bachan ko private school me padhana bahut pareshani hoti hai aise me kaise sambhav hai inki fits aur sugar ki private treatment ye sab kaise sambhav hoga app hi hamara margdarshan kijiye koi bhi sakt niyam mat lijiyega bas aapse itni prarthana haRespected sir,
I am an architecture student and and I believe architects play a major role in shaping up any society, city or even a nation. I believe that architects should also be involved with your team at the central level in policy making as well as building a great nation. After the independence, now changes have begun to happen, and we can change the way world will look towards new rising India. From railway stations to streets, and from street to nation we can highlight India globally.Dear sir,
सभी सांसद और विधायकों के लिए आईएएस और आईपीएस की तरह लिखित परिक्षा होना अति आवश्यक ह क्योकि देश के लिए असल नीति निर्धारक तो ये ही ह ,आखिर कम से कम इन को सामान्य विज्ञान (Science) ,पर्यावरण,स्वास्थ्य,सामान्य अर्थशास्त्र,देश विदेश और देश के संविधान के बारे में तो पता होना चाहिएwhy the first class is mandatory in all central govt jobs and other students are not eligible like 2nd class and pass candidates please solve this type of criteria all are talented even 2nd class or pass class people and sometime health problems or else some rural people representation problem in exams but everyone has its own hidden talent so please please terminate this type of criteria of job selection and give the way to young and talented people to move the india very well and smartDear sir shri,
मोदी जी देश में विज्ञान और शिक्षा को बढ़ावा दिए जाने के जरूरत ह और अंधविस्वास रोकने कीRespected PM,
Please i request you to speak about Road Safety, Safety on work and Home safety for women's in the Mann Ki Baat on 26th March.Our Beloved Prime Minister, I'm an start up entrepreneur and trying to establish a company to do research and marketing of high quality and high yielding seeds suitable for changing climate. From November month sales have reduced due to scarcity of money in the market as well a percieved fund shortage. The pain we are suffering is multiple, like growing interest burden, sales loss, employee related expenses, slow recovery of outstanding in the market. Now we are at the verge of collapse. Bless.Mother Cow is NOT an animal for 80 crore Hindus since the beginning of Era.Mother Cow is savior of our life.We demand that Mother Cow should be declared as "RASHTRA MATA" in our Constitution and there should be a separate ministry(Gau Mantralay).• We are importing Idols of Mother Cow, and its server Lord Krishna from various countries and shamelessly exporting Beef(Meat of Cow) to other countries just to gain foreign currency which is shameful.#Indian atheles grab 18 medals in winter paralympics 2017# We are proud of them,we must recognize their efforts and hard work and they must be rewarded appropriately. Tejvir Singh ,promoted to class 8th.respected dearest prime minister,
I am proud for your supreme decesion to upgrade to afflux to my country. the mission about house in all has been overwelmed me . I pray to you to which can be able to get the opportunity of your plan in fobour of me . Bank denied to give housing loan to me due to my agree culturing earning .
my request is kindly to make an arrangement to give housing loan foran agriculturist and oblige .Respected Prime Minister,
Thanks for taking India toward GST. My maan Ki Baat is that , in GST all suppliers of Goods And Services will get Compliance Rating from GSTN for proper return filling loss mismatch and prompt tax payment My suggestion is that this rating should be used at par

Plain Text Tab Width: 8 Ln 45, Col 1 INS

Nltk implementation

- NLTK is a platform for building Python programs to work with human language data.
- It provides easy-to-use interfaces, such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning etc.
- NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”
- Natural Language Processing with Python provides a introduction to programming for language processing. It guides one through the fundamentals of writing programs, categorizing text, analyzing linguistic structure, and more.

Code to separate noun from a .txt file

```
import nltk
import csv
import sys
reload (sys)
sys.setdefaultencoding('utf8')
#f=open('n.txt','r')

with open('n1.txt','r') as f:

        for line in f.readlines():


                tokens = nltk.word_tokenize(line)
                tagged = nltk.pos_tag(tokens)
                nouns = [word for word,pos in tagged \
                if (pos == 'NN' or pos == 'NNP' or pos == 'NNS' or pos== 'NNPS
                downcased = [x.lower() for x in nouns]
                joined = " ".join(downcased).encode('utf-8')
                into_string = str(nouns)

                output = open("m.txt", "a")
                output.write(joined)
                output.write("\n")
output.close()
```

Output

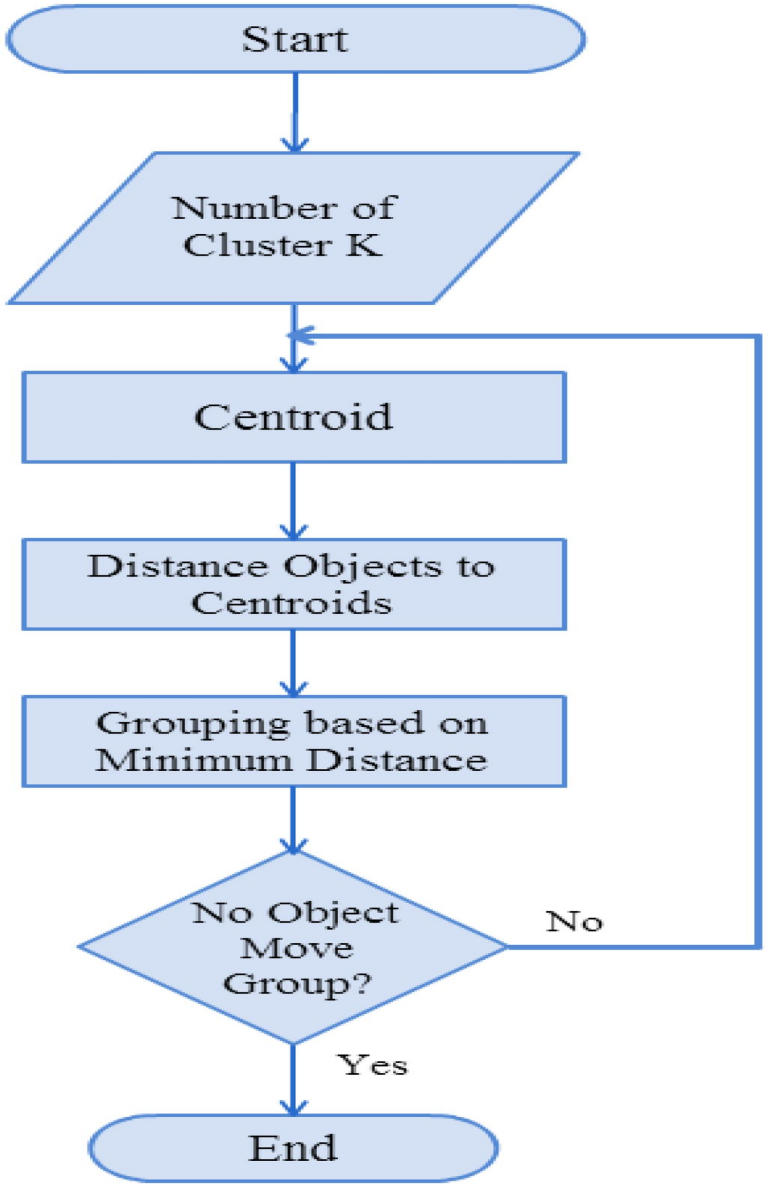
nouns.txt (~/artificialintelligence) - gedit

```
issue
thank
regard

chowk
value traction issues app issues city people pavements barriers destruction property traffic lights/street lights request zebra
crossing taxi/auto stands garbage hawkers/kiosks dustbins.for bettervisibility trackingprogress force shoes manufacturing need
taxes % pan gutkha cigrattes tobacco products alcohol please ban alcohol families sir
village kirmach distt kurukshetra wala hu gaon ki panchayat ki taraf se dustbin lgwaye pr ne ke apne ghar rkh liye hai aise logo ko
sir ap hi smjha skte hai chahta hu ke ap topic pe mann ki kahe work organisation lakhs clothes people city attention project break
innovation peoples summit states target focus % industries investors % concept countries.bring technology tycoon plants materials
traffic cops traffic power money tickets/ challen vehicle no/insurance/dl area corruption focus issues turn people fine kind system
place selfless services hospital wheels

physician diabetologist hospital years
services physician diabetologist days year ‘lifeline train
consideration
cry problems water people towards techniques water water table things farmers water crops summers
cost automation towards water conservation
reluctance digitization offices hole process interaction govt officials medium office work lot
decision demonetization interest nation prohibition liquors health status family development level class country living country
person decision diverse country.namaskar
problems transport sector wastage time fuel money toll collection centers tags payment system situation sir
attention martyr hero comrades hero world birth death anniversary something heroes air please action
fool measure tool
industries % % product department product import target pf sale tax industries motive request people garbage segregation home.start
locality model example implement city.knowledge activity folks.unsure attendance door2door time phones dhan accounts now.banks
activity.prototype use schools year books.instead books books books pupil process revenue trees process marks books year
tak mai janta hu mitti hai jise bharat mata kehte hai bharti keh puja karte hai bi yhi mante hai
hamara hai naa fir iske kyu hai ham apne bharat mata bol hai par jai bole
tak ki pakistan ka vi sirf ek hai hai
ke bi ek hi hai
aai bi angreio ki dharohar ko kyu dho rhe hai.hello
```

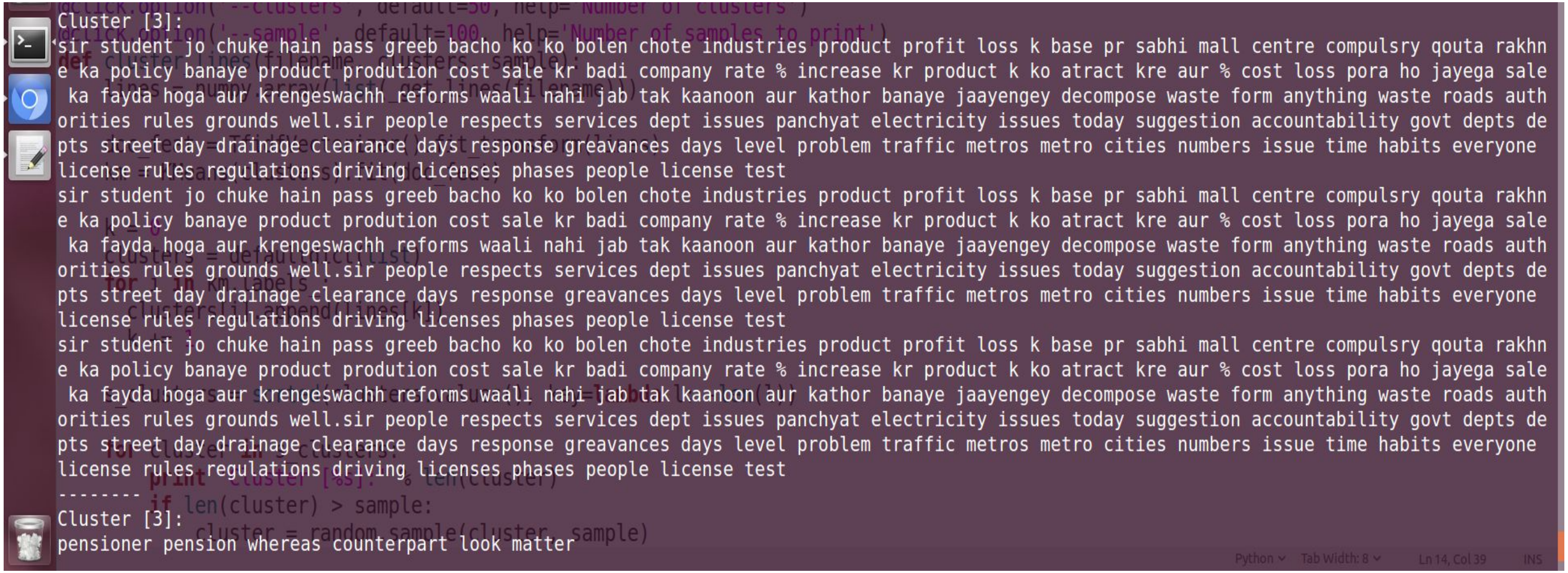
Plain Text Tab Width: 8 Ln 170, Col 125 INS

Application of clustering algorithm, i.e., KMeans

- ***k*-means clustering** is a method of vector quantization, used for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- A distance measure is needed to determine the "closeness" of instances.
- Classify an instance by finding its nearest neighbors and picking the most popular class among the neighbors.


Start
Number of Cluster K
Centroid
Distance Objects to Centroids
Grouping based on Minimum Distance
No Object Move Group?
No
Yes
End

Clusters formed after application of K Means Algorithm



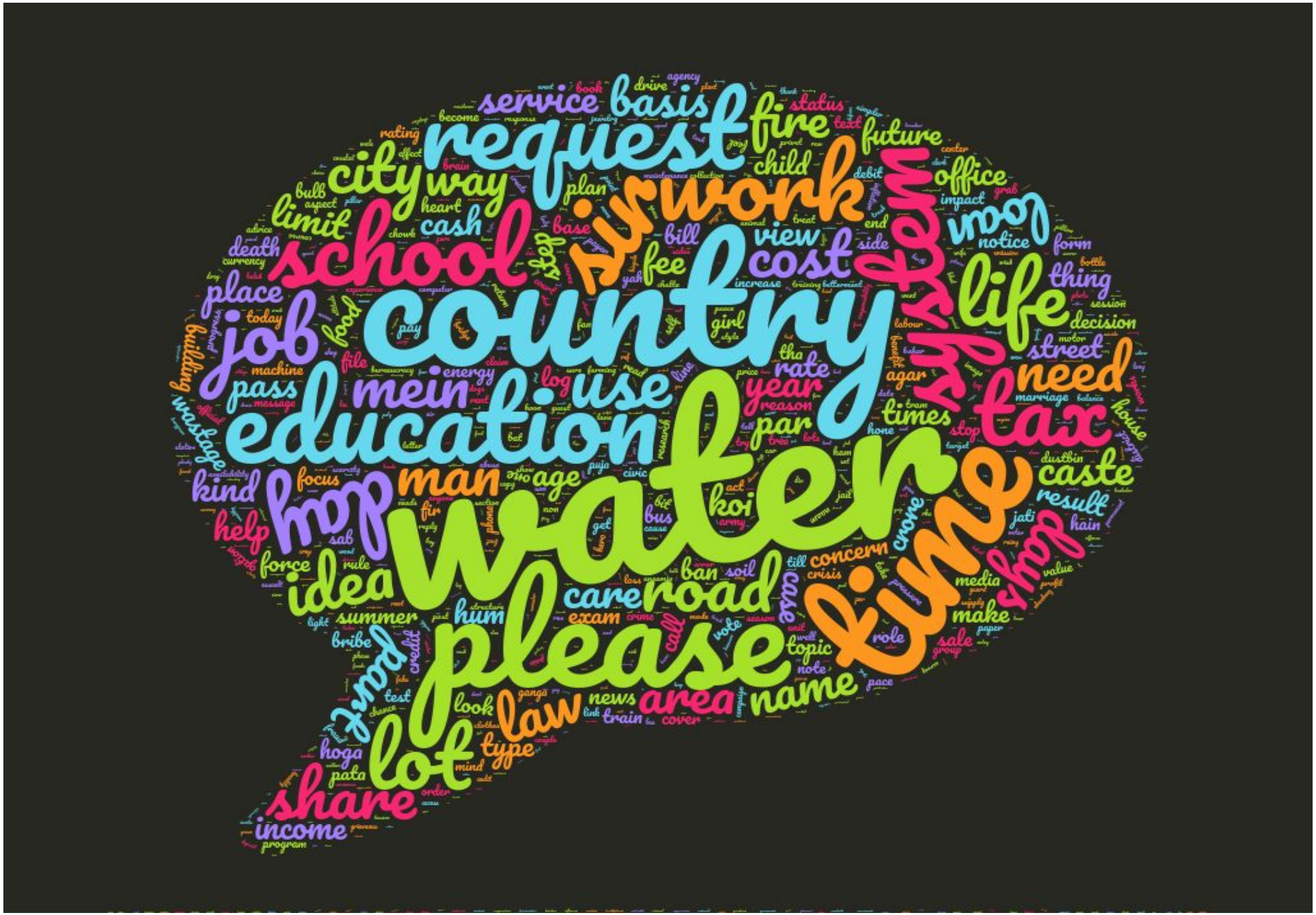
```
Cluster [3]:
sir student jo chuke hain pass greeb bacho ko ko bolen chote industries product profit loss k base pr sabhi mall centre compulsry qouta rakhn
e ka policy banaye product prodution cost sale kr badi company rate % increase kr product k ko atract kre aur % cost loss pora ho jayega sale
 ka fayda hoga aur krengeswachh reforms waali nahi jab tak kaanoon aur kathor banaye jaayengey decompose waste form anything waste roads auth
orities rules grounds well.sir people respects services dept issues panchyat electricity issues today suggestion accountability govt depts de
pts street day drainage clearance days response greavances days level problem traffic metros metro cities numbers issue time habits everyone
license rules regulations driving licenses phases people license test
sir student jo chuke hain pass greeb bacho ko ko bolen chote industries product profit loss k base pr sabhi mall centre compulsry qouta rakhn
e ka policy banaye product prodution cost sale kr badi company rate % increase kr product k ko atract kre aur % cost loss pora ho jayega sale
 ka fayda hoga aur krengeswachh reforms waali nahi jab tak kaanoon aur kathor banaye jaayengey decompose waste form anything waste roads auth
orities rules grounds well.sir people respects services dept issues panchyat electricity issues today suggestion accountability govt depts de
pts street day drainage clearance days response greavances days level problem traffic metros metro cities numbers issue time habits everyone
license rules regulations driving licenses phases people license test
sir student jo chuke hain pass greeb bacho ko ko bolen chote industries product profit loss k base pr sabhi mall centre compulsry qouta rakhn
e ka policy banaye product prodution cost sale kr badi company rate % increase kr product k ko atract kre aur % cost loss pora ho jayega sale
 ka fayda hoga aur krengeswachh reforms waali nahi jab tak kaanoon aur kathor banaye jaayengey decompose waste form anything waste roads auth
orities rules grounds well.sir people respects services dept issues panchyat electricity issues today suggestion accountability govt depts de
pts street day drainage clearance days response greavances days level problem traffic metros metro cities numbers issue time habits everyone
license rules regulations driving licenses phases people license test
--------
Cluster [3]:
pensioner pension whereas counterpart look matter
```

Segregating Data on the basis of likes

- The data file was extracted with the number of likes of each comment.
-
- To seperate the relevant comments from the data.
-
- The suggestions having greater than a fixed number of likes(used here is 10) were seperated out.

Where the likes came from?

```
<span>
        <a class="like_count" title="Login to Like"
href="
https://www.mygov.in/user/login?r=node%2F267721%3Ffie
ld_hashtags_tid%3D%26sort_by%3Dcreated%26sort_order%3
DDESC%2F%23comment-wrapper-10109651 1">Like</a>
                                <span
class="like_count_value">(21)</span>
                              </span>
```


water
country
education
please
time
request
work
school
worry
tax
job
life
need
mein
use
road
idea
lot
share
law
area
name
city
service
basis
fire
future
child
office
loan
way
limit
cash
cost
view
fee
bill
plan
place
pass
man
age
caste
result
days
care
kind
help
force
focus
koi
year
rate
par
times
street
decision
thing
notice
form
income
hoga
pata
type
look
news
train
cover
test
bribe
summer
exam
hum
topic
make
media
role
sale
concern
crisis
crore
jati
hain
house
marriage
stop
dustbin
wastage
building
machine
message
energy
file
food
pay
today
death
currency
advice
aspect
bulb
heart
base
step
rating
effect
become
response
book
drive
agency
status
text
maintenance
collection
impact
debit
end
side
increase
training
betterment
session
labour
motor
agar
tram
hone
ham
tree
lots
reason
research
civic
puja
act
bus
get
cause
army
ban
soil
till
case
unit
vote
call
crime
well
note
pace
group
paper
value
profit
supply
jail
target
order
program
mind
clothes
couple
link
light
credit
rule
sab
fir
phone
non
section
reply
scarcity
availability
official
station
letter
girl
farming
price
line
read
claim
bureaucracy
log
rent
guest
gift
peace
style
chalte
yah
computer
experience
chowk
brain
pillar
coastal
jewelry
plant
thank
singular
center
dark
grab
bottle
balance
benefit
image
pressure
season
crore
organic
loss
structure
post
guard
treat
journey

Presented By:

Ankita Jain(001)
Aishwarya Gupta(035)
Anamika Katyal(039)
Nitika Jain(040)
Priya Bansal(075)
Prerna Killedar(076)