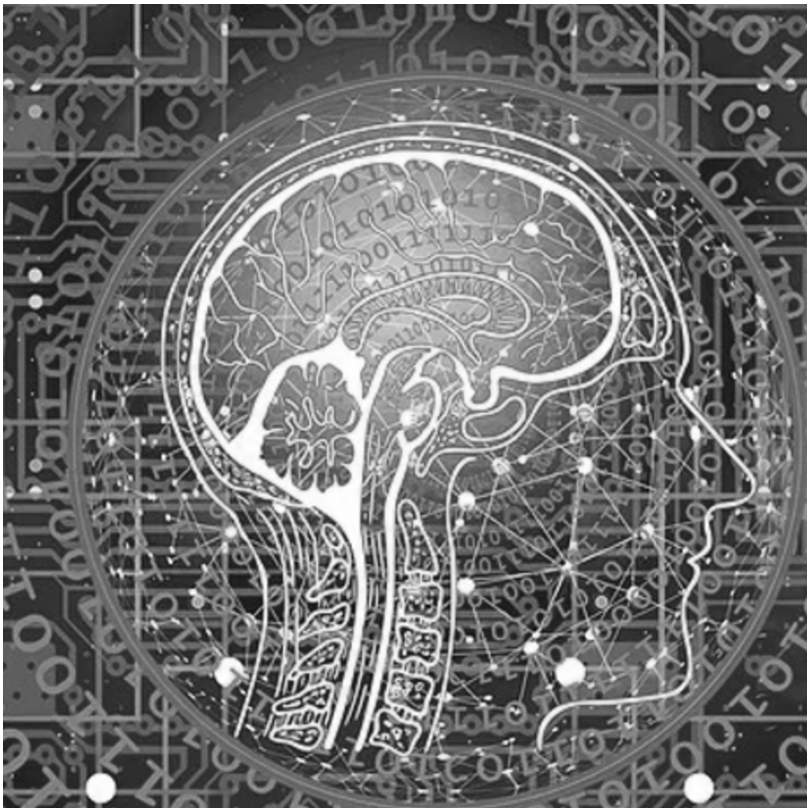


Coursera IBM-- Unsupervised Machine Learning: Clustering Final Project

By: Liam Webster

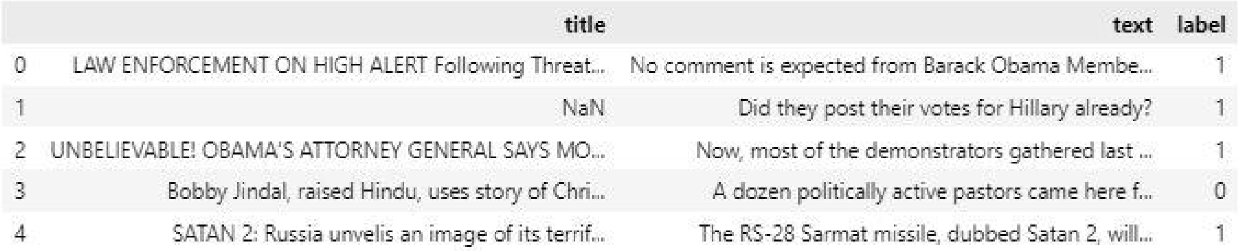
7/27/2022

Objective:

The main objective of this analysis was to cluster fake and real news articles and explore the differences in properties of real vs fake articles. Specifically the goal was to engineer a definitive two cluster model which accurately created a robust boundary between real and fake news articles. This model could then be used in applications in which filtering out fake articles is important.

Dataset:

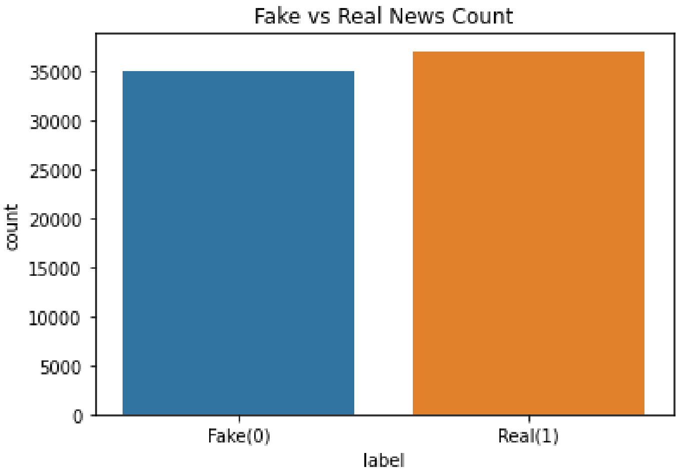
The dataset used in this analysis contains 72 thousand rows and four columns. Three feature columns– “Index”, “Title”, and “Text” – containing the index, the title of the news article, and a keyword summary of the article, respectively. The last column being “label” which indicates whether the article is fake or real. The dataset is a combination of 4 other datasets thus amassing a total of 72 thousand observations.

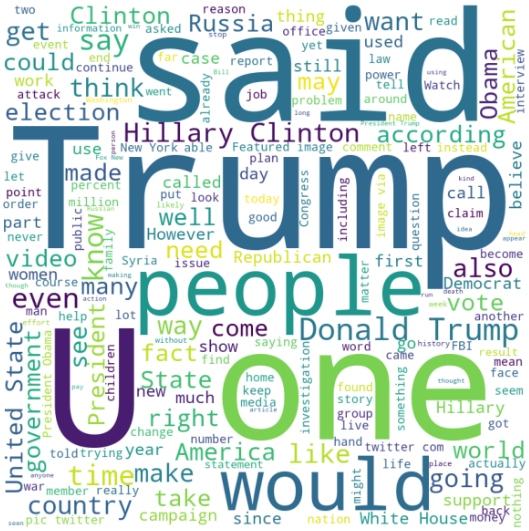


	title	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1	NaN	Did they post their votes for Hillary already?	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1

EDA:

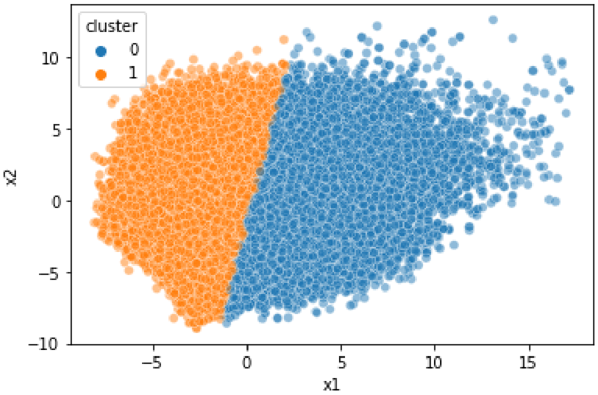
The first step taken was a manual inspection of the .csv file. The dataset was checked for Null values of which were removed. The “text” and “title” columns were combined into one “text” column. After some more brief data cleaning a couple general plots were created. Including a plot displaying the count of fake vs real articles. The plot showed that the dataset was evenly distributed. Word clouds were created displaying common words amongst the real and fake articles.

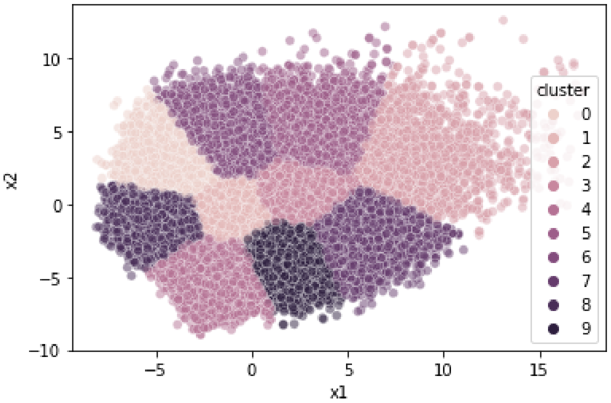




Models:

The first model used was a KMeans model with two clusters. The “text” column was put through a filter and lemmatizer to create a corpus. This corpus was then vectorized creating a TF-IDF matrix. This matrix was then scaled and fit to a KMeans two cluster model. This model was cross referenced with the dataset’s “label” column, and it was found to cluster 70% of the articles correctly. This model was further explored using principal component analysis. PCA was performed on the TF-IDF matrix. Instead of the original 1000 column TF-IDF matrix PCA was able to simplify the matrix to just 2 components while maintaining 70% accuracy. Common descriptive words were found for each respective cluster. The next model explored was facilitated by creating a cluster count elbow curve. This plot showed 10 clusterers to be the optimal choice for efficiently minimizing the inertia of the clusters. From this a 10 cluster KMeans model was fit to the PCA data. With this model much was to explore. First a plot of the cluster was created. Next a dive into the relation between each of the ten clusters and their real vs fake article distribution. It was found that articles within cluster 2 and cluster 9 had an estimated 10 percent chance of being fake while articles within cluster 4 and 6 had an estimated 89 percent chance of being fake, Common words within each of clusters were found.





Key Model:

After analysis of all the models, it was found that the KMeans model with 10 clusters produced the best fit for application. With this model future articles could be clustered with the model, of which an estimation could be made whether the article was fake or real.This model produced clusters with the heaviest weight(most observations) having the highest predictability. While clusters with lower weight(least observations) had lower predictability. Thus while in certain cases future clustered articles may have an undetermined prediction but most will have a high factor of predictability.

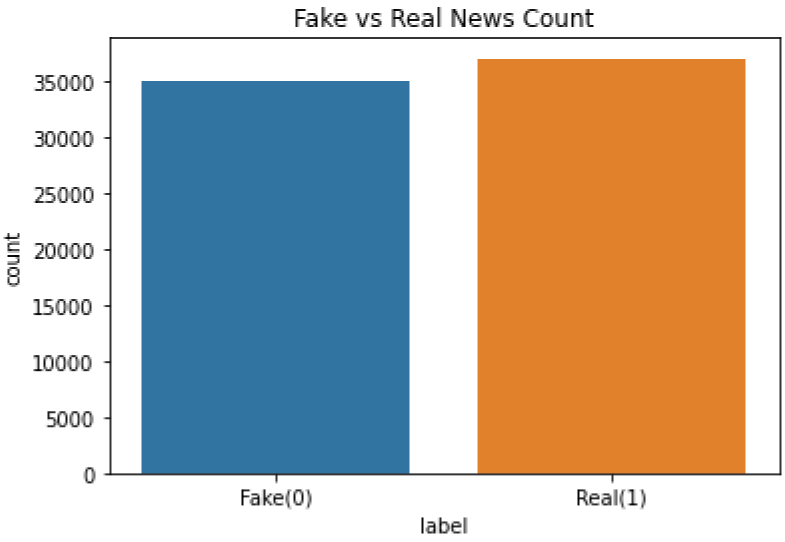
Findings and Insights:

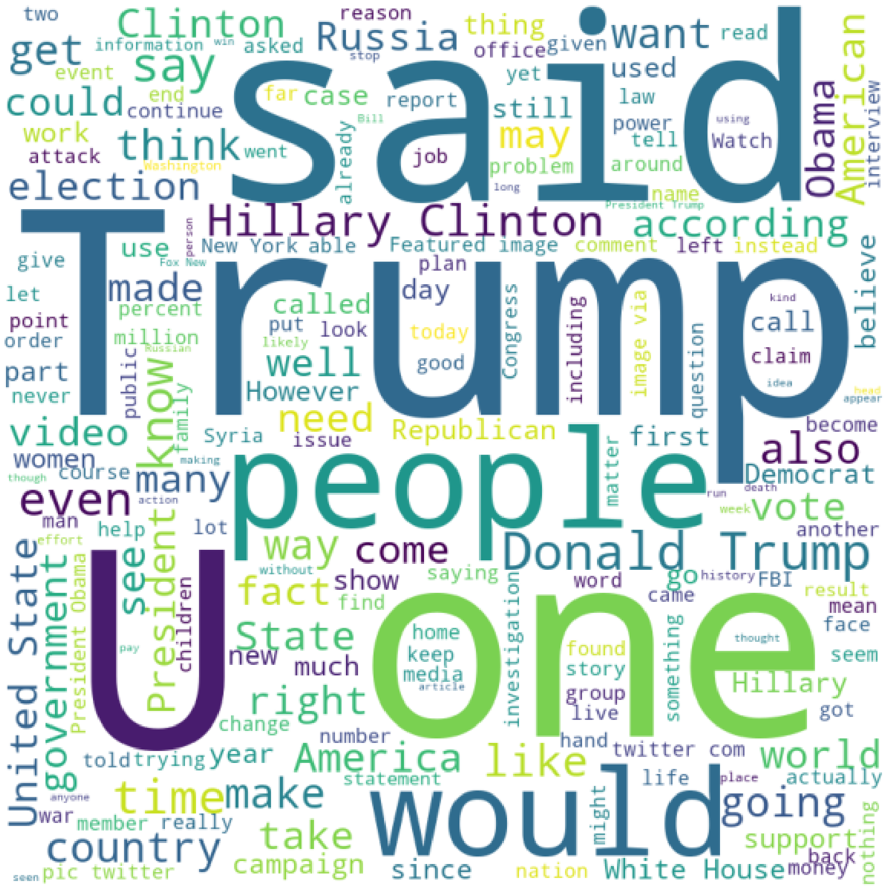
Working with large data requires a lot of computing power and patience. Creating the data corpus from the 70 thousand articles took over two hours. Words such as ‘Trump’, ‘President’, and ‘people’ were commonly found in both real and fake articles. While words such as ‘Clinton’, ‘think’, and ‘reality’ were commonly found in real articles and words such as ‘Government’, ‘illegal’, and ‘Reuters’ were commonly found in fake articles.

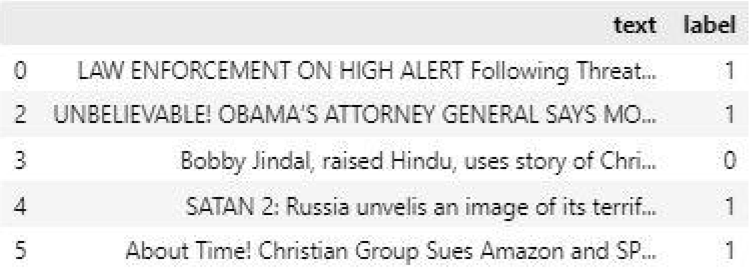
Future:

The data could be better cleaned and filtered. Creating a clearer and more definite corpus which would help create more defined cluster boundaries. Possibly a new analysis of just the original “title” or “text” column, before the merge, could present new findings. More sophisticated natural language processing techniques could be performed.

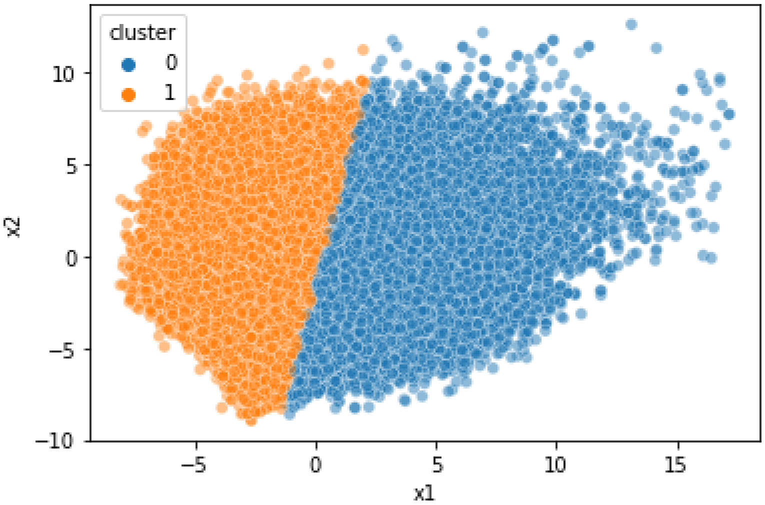
Appendix:

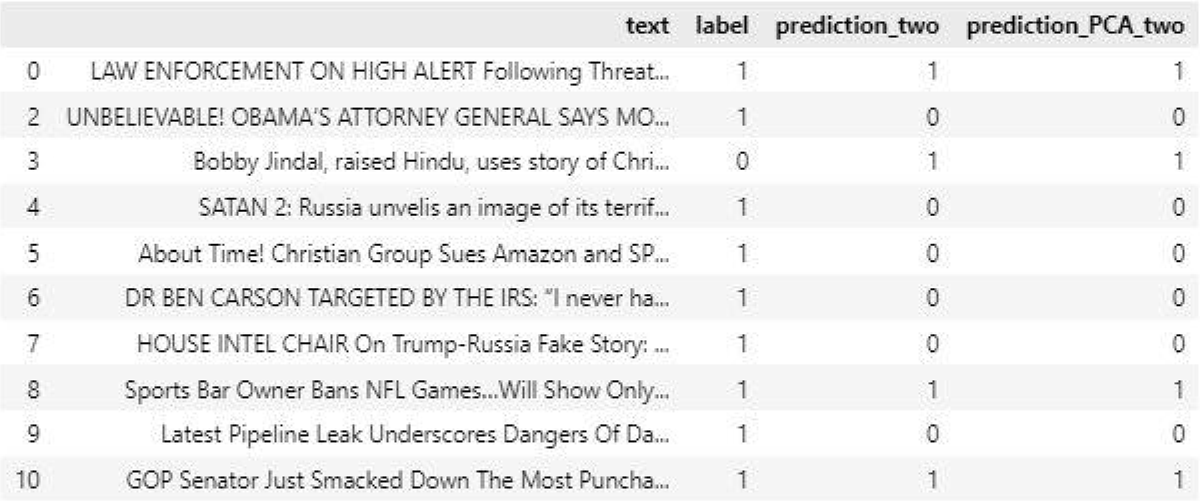

Fake vs Real News Count
count
35000
30000
25000
20000
15000
10000
5000
0
Fake(0)
Real(1)
label


said
Trump
one
people
U
would
Donald Trump
Hillary Clinton
United State
President
Clinton
Russia
American
Obama
want
get
say
could
think
election
according
know
government
America
like
time
make
country
world
going
also
Democrat
vote
video
even
many
come
way
fact
right
State
need
Republican
well
made
called
thing
may
still
used
case
report
first
take
campaign
support
White House
since
year
see
two
reason
information
asked
office
given
read
event
end
continue
far
yet
law
power
tell
Watch
interview
work
attack
went
already
job
problem
around
name
Washington
use
New York
able
Featured image
comment
left
instead
give
plan
let
day
percent
put
look
today
Congress
including
image via
question
believe
point
million
call
claim
order
likely
good
part
public
family
However
never
Syria
issue
become
women
course
matter
another
man
lot
help
show
saying
go
history
FBI
result
mean
face
home
keep
media
found
story
group
live
Hillary
seem
got
new
much
children
investigation
something
change
number
hand
told
trying
twitter com
statement
life
actually
member
really
war
might
pic twitter
nation
back
money
nothing
President Obama
effort
find
came
word
thought
article
idea
kind
person
Fox New
Russian
making
though
action
without
pay
anyone
seen
place
week
run
death
head
appear
stop
win
Bill
using
long
President Trump

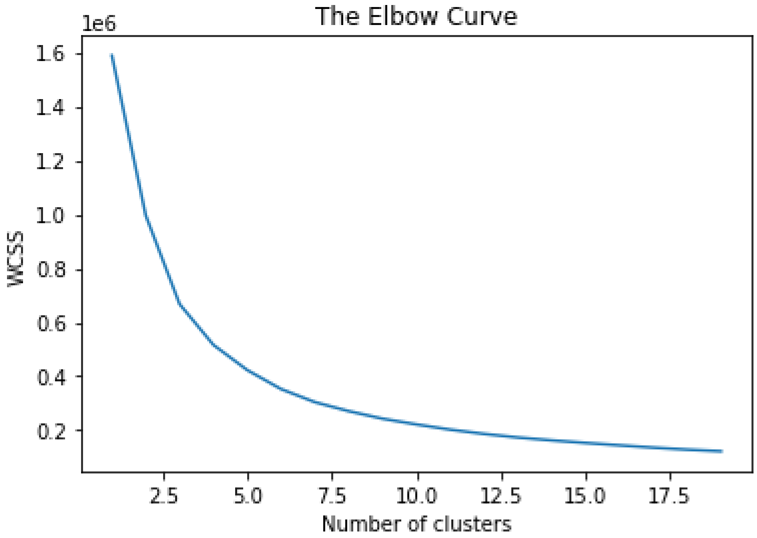


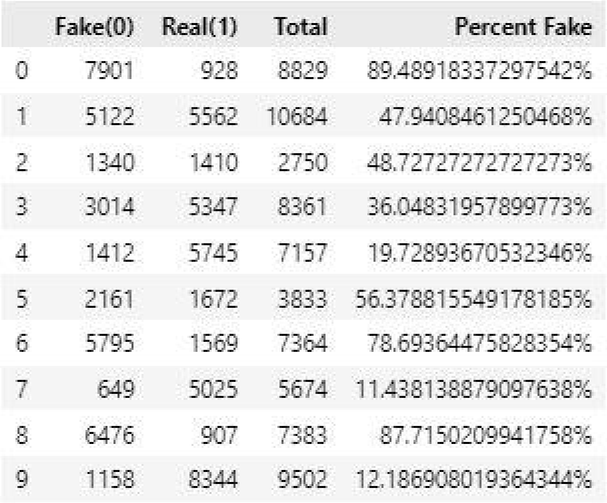
	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	0
4	SATAN 2: Russia unvelis an image of its terrif...	1
5	About Time! Christian Group Sues Amazon and SP...	1



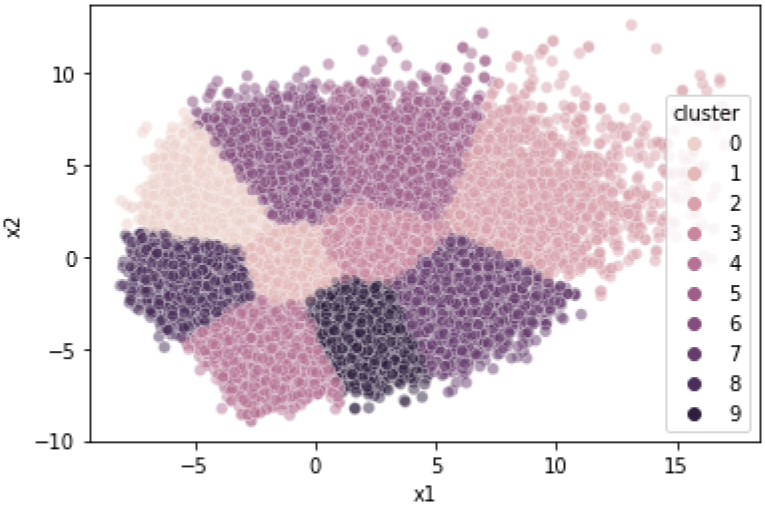


	text	label	prediction_two	prediction_PCA_two
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1	1	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1	0	0
3	Bobby Jindal, raised Hindu, uses story of Chri...	0	1	1
4	SATAN 2: Russia unvelis an image of its terrif...	1	0	0
5	About Time! Christian Group Sues Amazon and SP...	1	0	0
6	DR BEN CARSON TARGETED BY THE IRS: "I never ha...	1	0	0
7	HOUSE INTEL CHAIR On Trump-Russia Fake Story: ...	1	0	0
8	Sports Bar Owner Bans NFL Games...Will Show Only...	1	1	1
9	Latest Pipeline Leak Underscores Dangers Of Da...	1	0	0
10	GOP Senator Just Smacked Down The Most Puncha...	1	1	1





	Fake(0)	Real(1)	Total	Percent Fake
0	7901	928	8829	89.48918337297542%
1	5122	5562	10684	47.9408461250468%
2	1340	1410	2750	48.72727272727273%
3	3014	5347	8361	36.04831957899773%
4	1412	5745	7157	19.72893670532346%
5	2161	1672	3833	56.378815549178185%
6	5795	1569	7364	78.69364475828354%
7	649	5025	5674	11.438138879097638%
8	6476	907	7383	87.7150209941758%
9	1158	8344	9502	12.186908019364344%


x2
x1
cluster