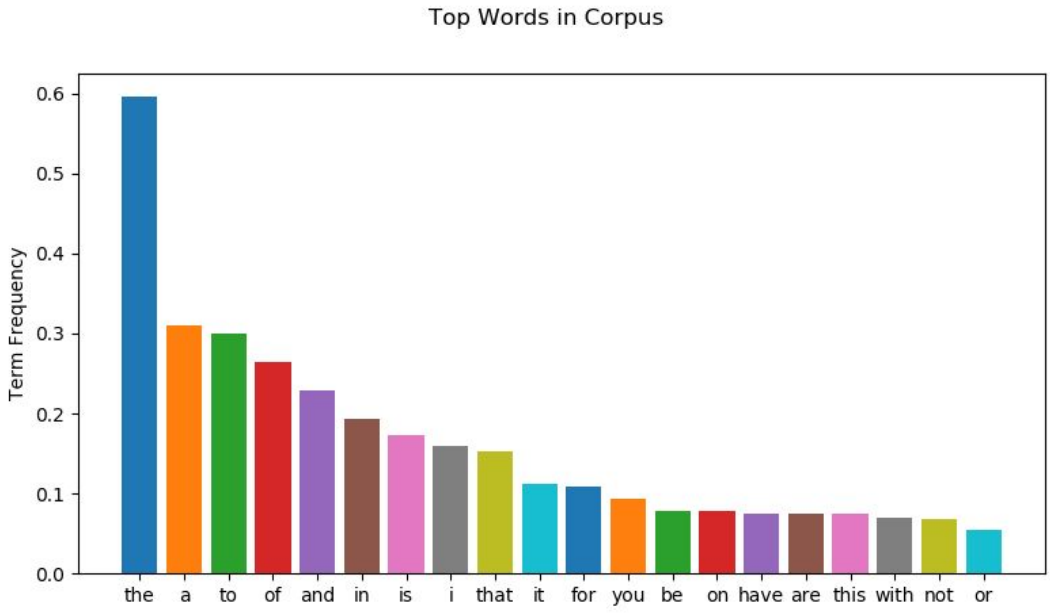
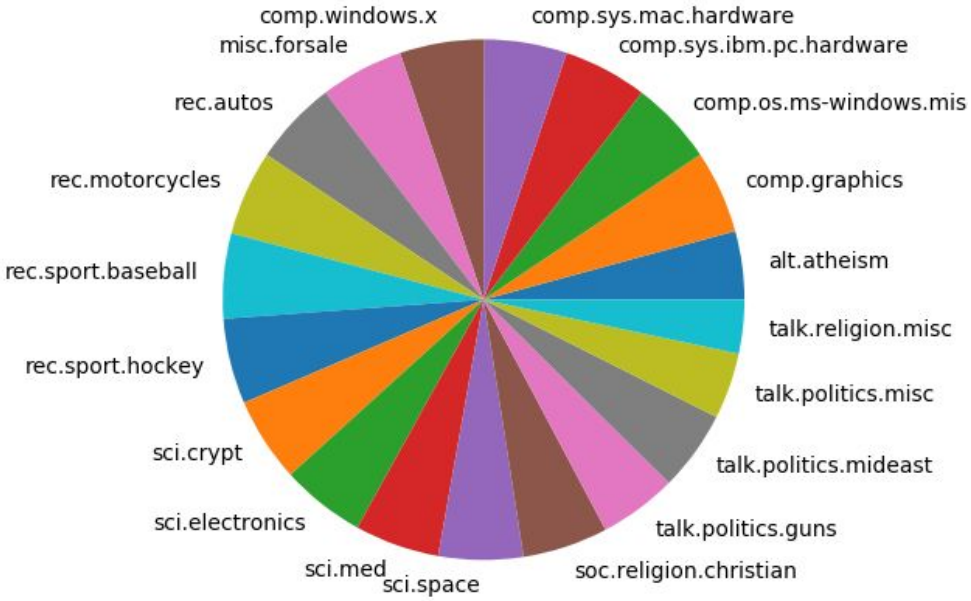
Jack Hirsh, Matt Hoffman, & Daniel Silver (Group 14)
Data Visualization Midterm Project Visualizations

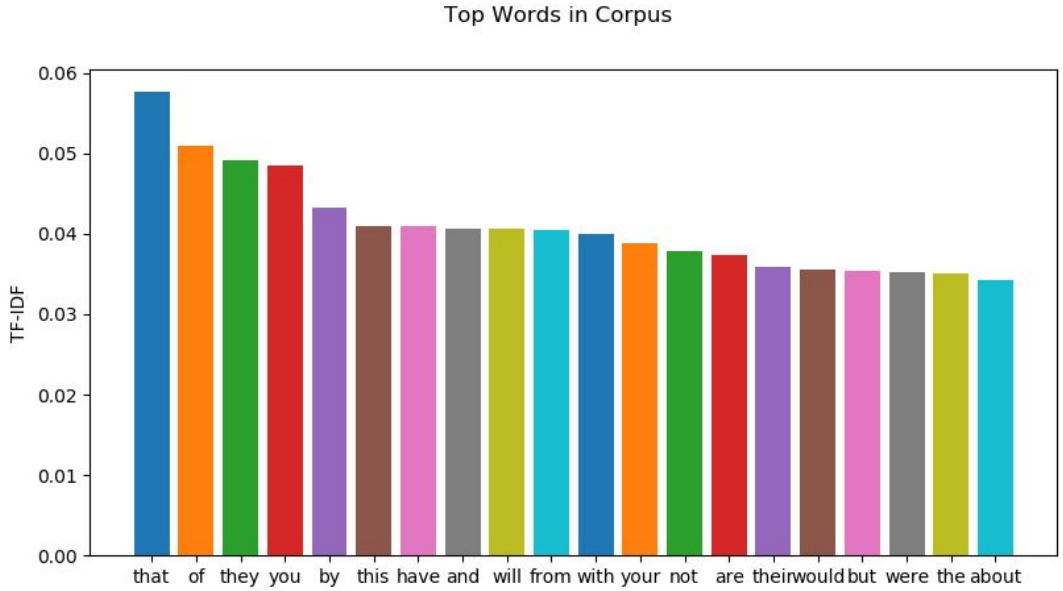
1. Visualize Statistical Information:

```
# of Documents:11314
# of Total Words:3406593
# of Unique Words:224662
```

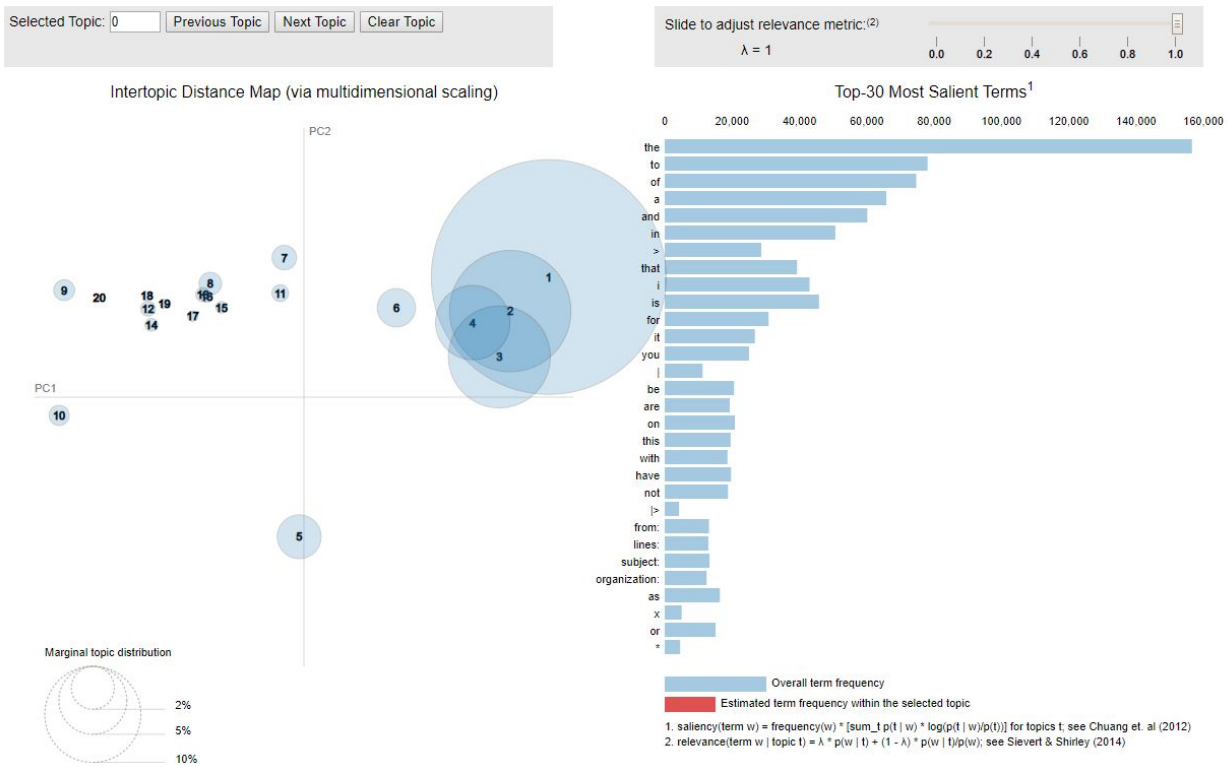




2. Build two different vocabularies upon different preprocessing ways; Learn Bag-of-words (BoW) and TF-IDF model with each vocabulary accordingly.

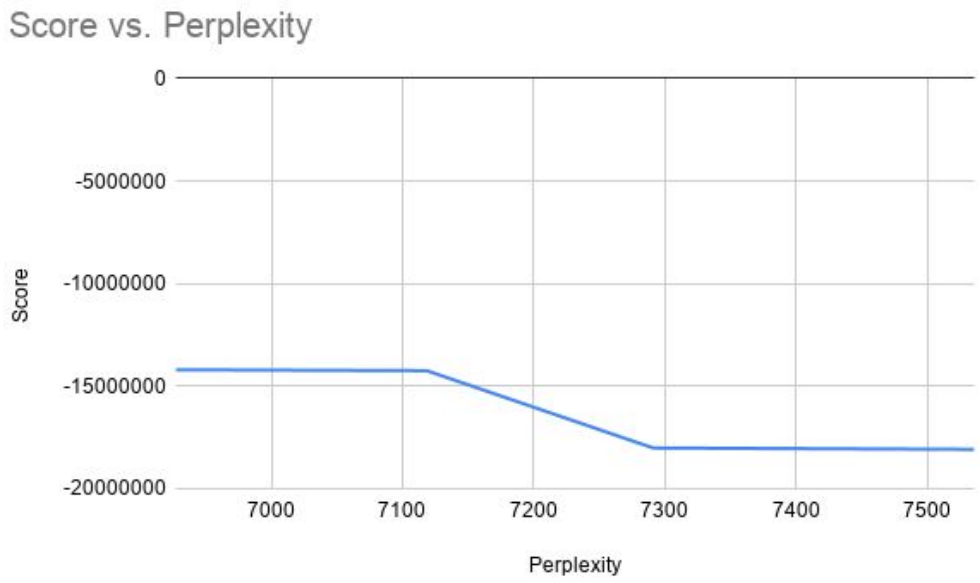


3. Visualize LDA for 2 vocabularies



This is the visualization for the unfiltered data set; the html is in the package, as well as the filtered version.

This score versus complexity is the result of training 6 different LDA models, across both vocabularies and multiple topics and calculating the score and perplexity of the data.



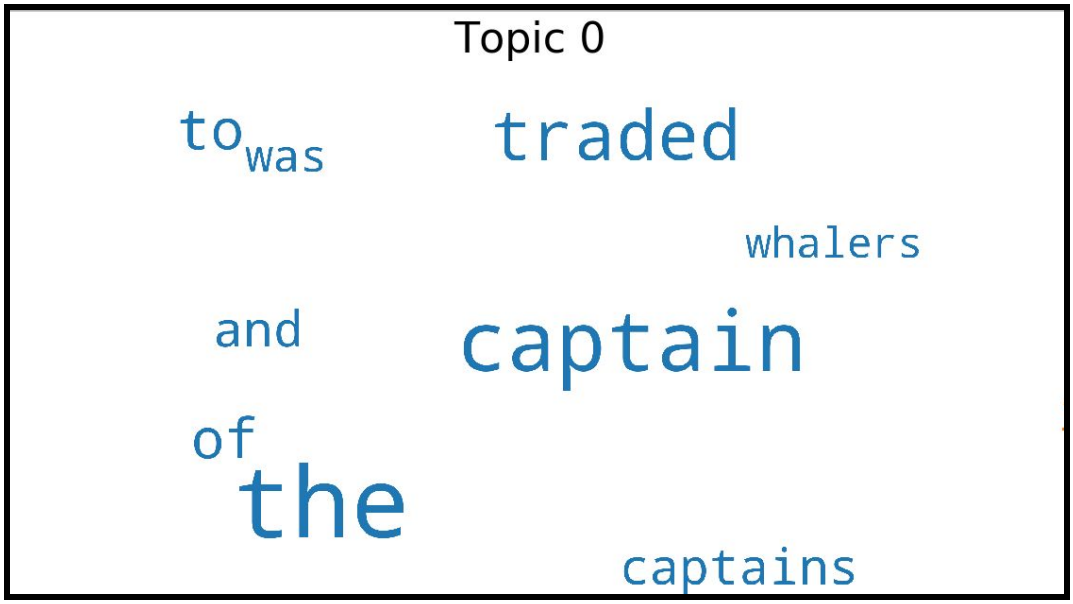
Top 10 words for each topic for filtered dataset (no headers or footers)

```
Topic: 0
output file use program time printf char entry input oname
Topic: 1
window widget application visual manager value xt alomar com widgets
Topic: 2
com men writes lib radar homosexual article gay don car
Topic: 3
turkish armenian people armenians edu turkey jews writes article greek
Topic: 4
edu windows com writes use article know like does just
Topic: 5
ax max g9v b8f a86 pl 145 1d9 0t 1t
Topic: 6
god people think does writes don like article human question
Topic: 7
government people key president law encryption mr right chip new
Topic: 8
people god edu just don know writes think say like
Topic: 9
file com writes don water entry power article like use
Topic: 10
00 10 25 15 11 12 20 14 16 13
Topic: 11
edu writes article com cs mouse 55 mit know subject
Topic: 12
drive scsi hard disk bus card drives controller like just
Topic: 13
cx w7 c_ uw t7 ck w1 chz lk hz
Topic: 14
writes article com edu don just people like gun know
Topic: 15
space edu nasa program software data information use graphics available
Topic: 16
edu writes team game article year think don like games
Topic: 17
db key use bit jpeg file ground wire des bits
Topic: 18
period pp pts power scorer play new good ny mv
Topic: 19
edu dos 00 writes article ah simms 34u air 7u
```

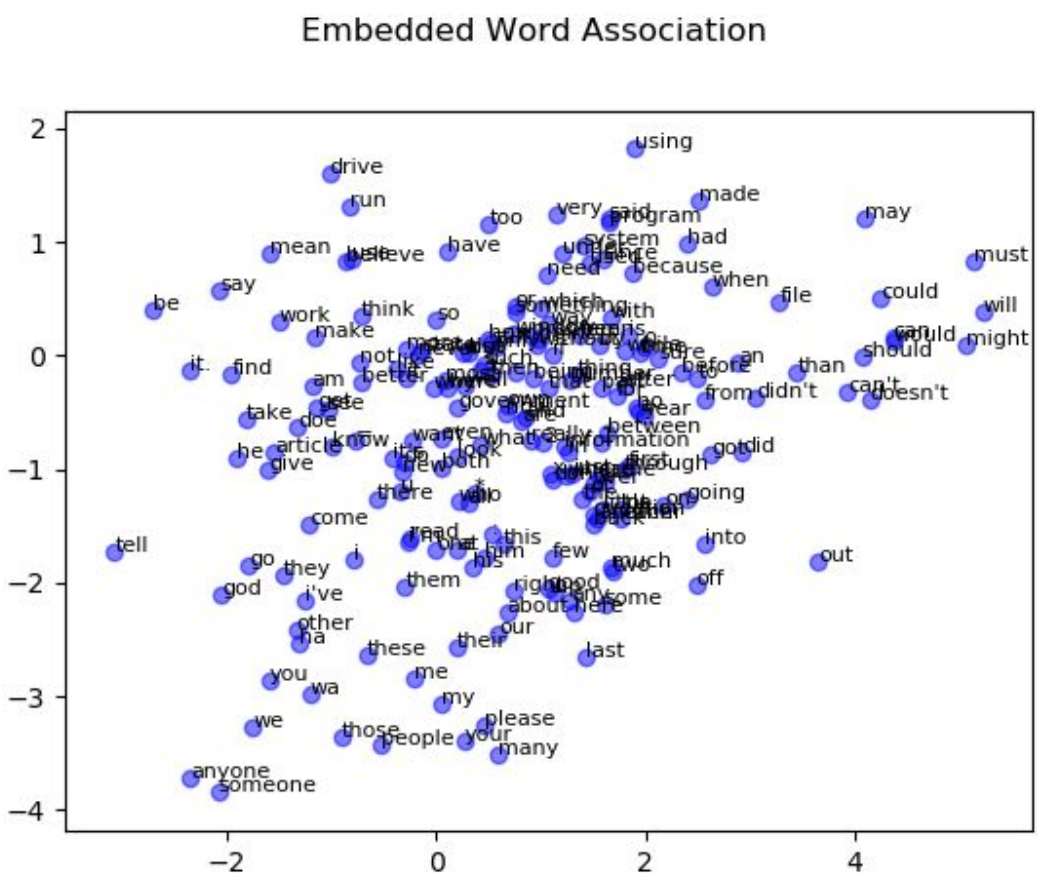
Top 10 words for each topic for original dataset

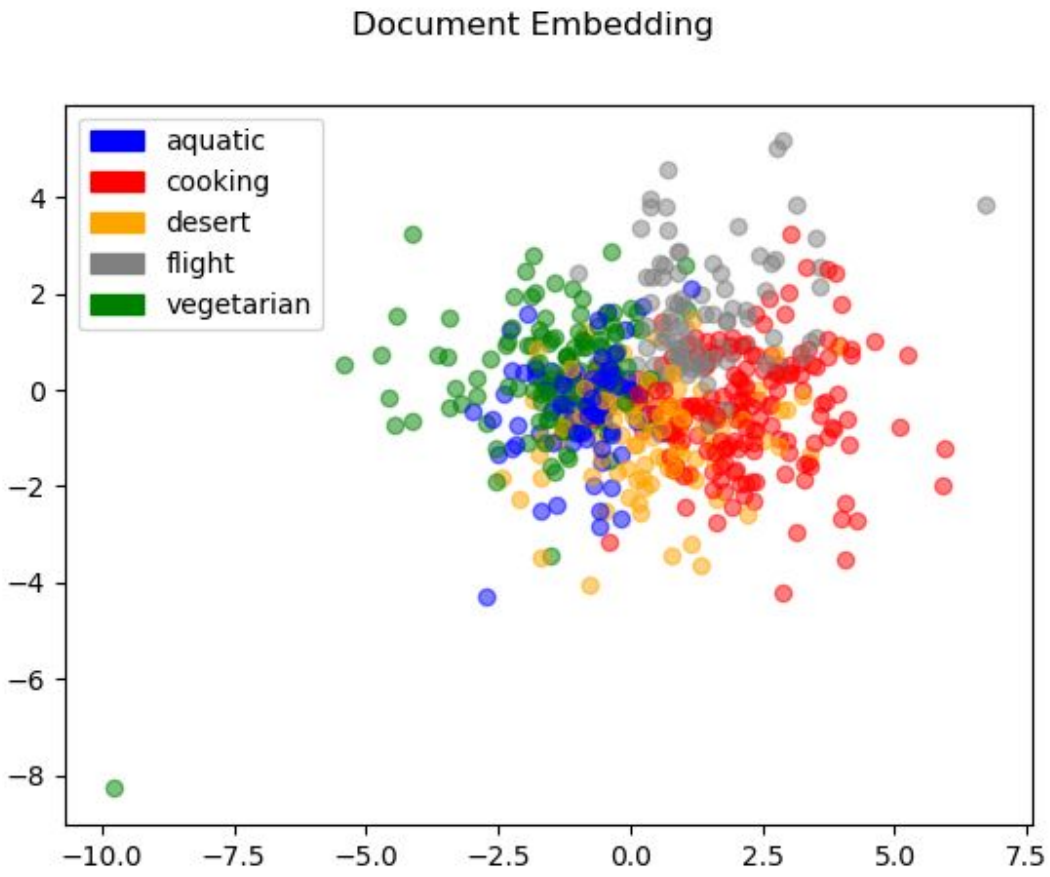
```
Topic: 0
edu lines subject organization bit card use color university com
Topic: 1
ax max g9v b8f a86 pl 145 1d9 0t 34u
Topic: 2
people said file armenian turkish know don armenians time new
Topic: 3
god edu people think don just subject believe jesus like
Topic: 4
edu com lines subject organization posting host nntp car writes
Topic: 5
0d mr stephanopoulos _o 145 president a86 6um 2di 34u
Topic: 6
edu writes subject organization lines article like com don people
Topic: 7
edu cx w7 c_ uw t7 17 ck chz lk
Topic: 8
edu ca subject lines organization writes article jesus just like
Topic: 9
edu com subject lines organization writes article god cs posting
Topic: 10
edu com writes organization subject article lines team don think
Topic: 11
edu subject lines organization com article posting university mail host
Topic: 12
edu writes article subject ca lines organization like don just
Topic: 13
people israel gun edu jews israeli subject writes think government
Topic: 14
00 10 25 15 11 12 20 16 14 13
Topic: 15
key space edu nasa encryption data information use clipper chip
Topic: 16
com edu lines subject organization scsi writes windows article host
Topic: 17
edu subject lines com organization windows use mouse article writes
Topic: 18
com subject edu lines organization motif server use widget available
Topic: 19
edu db lines sale subject organization university posting host nntp
```

Word Cloud of most common words in a single LDA topic:



4.

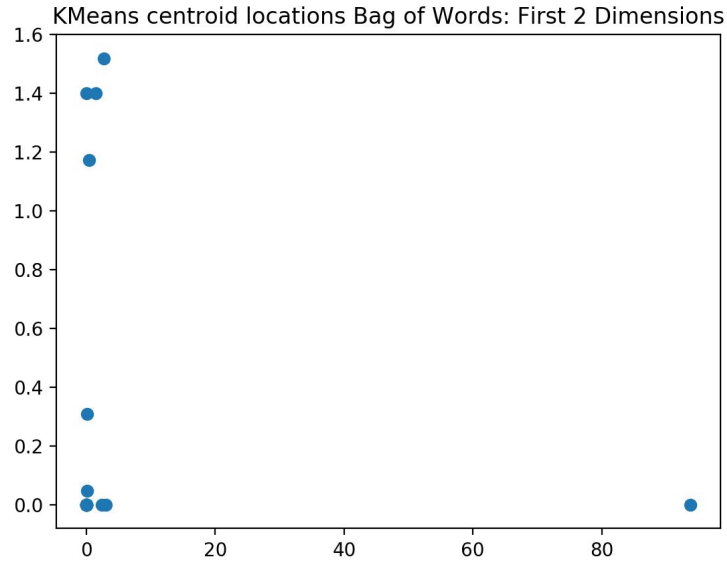

Embedded Word Association
using
drive
made
run
too
very
said
program
may
mean
have
system
had
must
believe
need
because
when
say
file
could
be
work
think
so
with
will
make
should
might
it.
find
am
than
didn't
can't
doesn't
take
between
give
got
did
going
come
tell
this
into
go
they
few
much
out
god
i've
them
off
other
about
here
our
these
their
last
you
me
wa
my
we
please
those
people
your
many
anyone
someone



5. K-means clustering (Mr. Tao said that our Naive Bayes' classifier could supplement K-means so we have included the results of it as well)

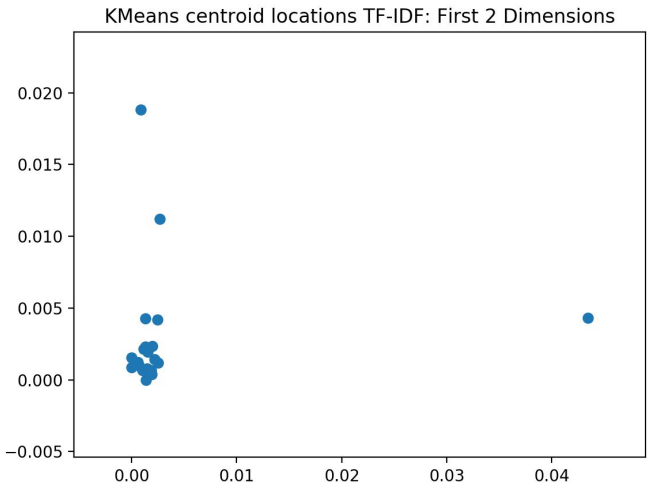
	Original Dataset	Filtered Dataset
Naive Bayes' TF-IDF Accuracy	82.90%	79.40%

a.) BoW




```
Homogeneity: 0.008
Completeness: 0.153
V-measure: 0.015
Adjusted Rand-Index: 0.000
Silhouette Coefficient: 0.150

Top terms per cluster:
Cluster 0: knowledge people god good religion bad provided food eat philosophy
Cluster 1: alexander stanford usa keywords ha offer ah sale university 415
Cluster 2: use sfasu suicide did ccsvax austin gas stephen account anonymous
Cluster 3: duo kind powerbook stanford memory does machine bit 68030 speed
Cluster 4: 25 people like time don space new use just right
Cluster 5: 16 place john doctrine die listserv genesis gov apr 25
Cluster 6: gun state said got pro people order bills 30 right
Cluster 7: af shuttle mil space elements orbital element current institute technology
Cluster 8: org john ii 000 state 17 tx keys territory cactus
Cluster 9: com article university posting don just host like nntp know
Cluster 10: initiative wiretapping bush ingres garrett urge com clinton strongly computer
Cluster 11: 42 alice com suggestions times article tiff hill wrote nj
Cluster 12: mcmaster ca board circuit western holly drive digital university wd
Cluster 13: stats skidmore 1993 nhl player final scores 1992 receiving season
Cluster 14: exe cis ufl usl dave figure work think tell just
Cluster 15: support xterm windows university program thanks don queensu ca like
Cluster 16: character font russ au spaces period displayed used sharp graphics
Cluster 17: jake useless com cs like sure article department stanford comes
Cluster 18: car sounds like know owner good just work bmw does
Cluster 19: gld columbia dare cc cunixb gary pat burns jets stanley
```

b.) TF-IDF



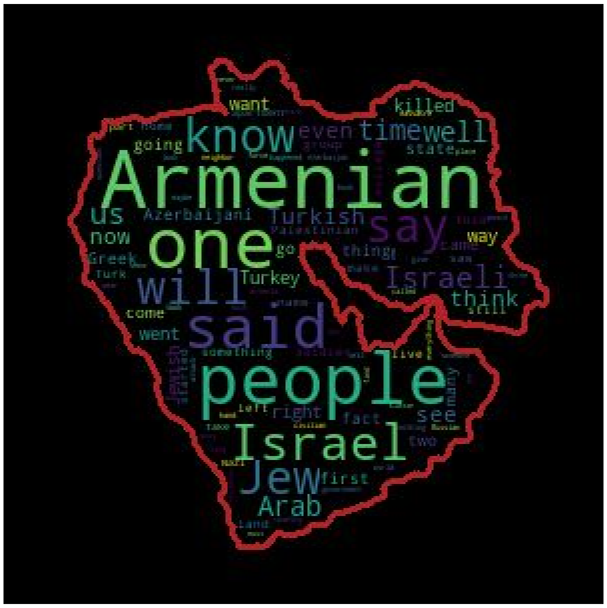

```
Homogeneity: 0.320
Completeness: 0.406
V-measure: 0.358
Adjusted Rand-Index: 0.123
Silhouette Coefficient: 0.008

Top terms per cluster:
Cluster 0: geb banks gordon pitt cs dsl n3jxp chastity cadre shameful
Cluster 1: god jesus people bible christian christians believe christ church say
Cluster 2: uga ai georgia covington mcovingt athens aisun3 michael programs 706
Cluster 3: clipper key chip encryption com keys escrow government netcom nsa
Cluster 4: scsi drive ide bus controller card isa eisa drives vlb
Cluster 5: ibm cc utexas austin com columbia drive purdue university posting
Cluster 6: sandvik apple kent newton com alink ksand cookamunga tourist bureau
Cluster 7: com article sun netcom bike posting nntp host ca like
Cluster 8: uk ac cs file university graphics thanks files program ftp
Cluster 9: gun stratus fbi batf com people guns koresh sw government
Cluster 10: game team hockey games ca players year baseball season play
Cluster 11: israel israeli jews arab jake arabs jewish uci israelis bony1
Cluster 12: virginia cramer optilink clayton gay men study university com homosexual
Cluster 13: windows window dos 00 mouse com ms file program problem
Cluster 14: mit lcs xpert expo athena enterpoop internet ai host nntp
Cluster 15: usc wpi angeles southern los aludra california ca worcester posting
Cluster 16: space nasa henry gov access digex alaska toronto pat shuttle
Cluster 17: university posting host nntp article just know like state ca
Cluster 18: hp com hewlett packard col cv fc tin newsreader ham
Cluster 19: caltech keith livesey sgi solntze wpd jon cco schneider pasadena
```

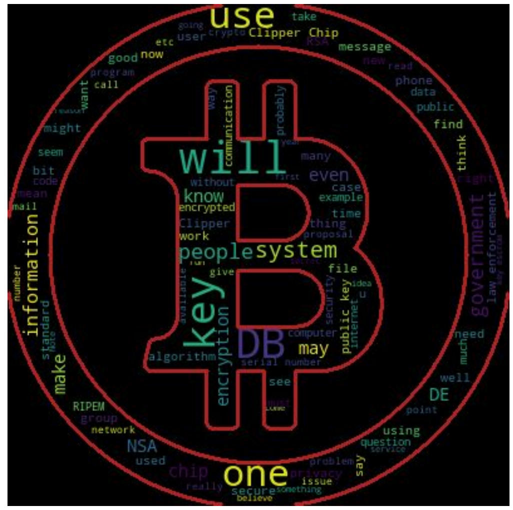
c.) Topics Distribution

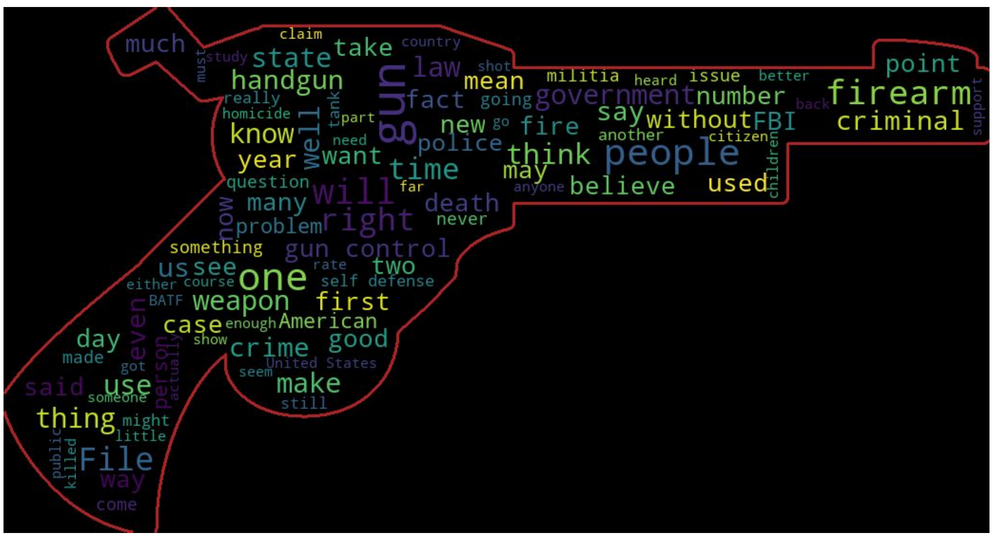
d.) Doc2Vec

7. Bonus Visualizations

Word clouds for individual categories contained within shapes that outline the main topic (Jupyter notebook used to generate word clouds in real time is included)


Armenian
people
said
one
know
will
say
Israel
Jew
Israeli
Arab
Turkish
Turkey
Azerbaijani
time
well
even
killed
think
us
now
want
going
way
came
went
come
fact
right
see
two
first
Greek
Turk
live
still


use
will
key
DB
one
system
people
know
encryption
information
government
law enforcement
Clipper Chip
message
phone
data
public
find
think
even
case
example
time
thing
proposal
file
may
algorithm
serial number
see
NSA
chip
privacy
secure
issue
say
well
DE
need
RIPEM
group
network
used
make
want
call
program
good
now
user
etc
take
might
seem
bit
code
mail
number
standard
work
encrypted
without
probably
communication
computer
public key
internet
security
using
question
point


much
state
take
claim
country
law
mean
militia
heard
issue
better
point
firearm
criminal
handgun
gun
fact
going
government
number
FBI
say
without
know
year
well
want
new
fire
police
think
people
time
may
believe
used
will
many
right
death
never
now
problem
question
something
gun control
us
see
one
two
self defense
weapon
first
case
American
day
crime
good
even
person
said
use
make
still
thing
might
little
File
way
come
public
killed
made
BATF
either
course
really
homicide
tank
part
need
shot
support
another
citizen
children
anyone
far
rate
enough
show
United States
seem
someone
actually
got